

Analyse d'un jeu de données réel

Fabien Haury

2022-04-30

- 1 Librairies
- 2 Données manquantes
 - 2.1 Jeu de données PhD_v1
 - 2.2 Jeu de données PhD_v2
 - 2.3 Visualisation des données manquantes
- 3 Problèmes
 - 3.1 Problèmes soutenances
 - 3.2 Problème homonyme Cecile Martin
- 4 Outliers
 - 4.1 Outliers director
- 5 Résultats préliminaires
 - 5.1 Langues
- 6 SQL
- 7 Travail en bonus
 - 7.1 Heatmap des données manquantes
 - 7.2 Problème Genre/Discipline/Langue
 - 7.3 Webscraping

1 Librairies

```
library(plyr)
library(tidyverse)
library(lubridate)
library(naniar)
library(scales)
library(ggsci)
library(GGally)
library(ggthemes)
library(lemon)
library(rvest)
```

`plyr` pour appliquer le paradigme «split-apply-combine».

`tidyverse` contient `tidyr`, `dplyr` et `ggplot2` pour la manipulation et visualisation des données.

`naniar` pour visualiser les données manquantes.

`lubridate` pour la gestion des formats de date.

`scales` pour modifier les formats d'échelles des graphiques.

`ggsci` pour différents thèmes de graphique.

`ggthemes` donne accès à des thèmes supplémentaires pour `ggplot2`

`lemon` donne accès à des fonction supplémentaires pour changer subtilement des aspects de `ggplot2`

`rvest` pour le web scraping.

2 Données manquantes

2.1 Jeu de données PhD_v1

2.1.1 Import du jeu de données

```
these_v1 <- read_csv("jeux_de_donnees/PhD_v1.csv")
head(these_v1)
```

```
## # A tibble: 6 x 1
##   `Auteur;Identifiant auteur;Titre;Directeur de these;Directeur de these (nom ~`
##   <chr>
## 1 "Saeed Al marri;;Le credit documentaire et l'onopposabilite des exceptions;Ph~
## 2 "Andrea Ramazzotti;174423705;Application de la PGD a la resolution de problem~
## 3 "OLIVIER BODENREIDER;;Conception d'un outil informatique d'etude des cinetiqu~
## 4 "Emmanuel Porte;;Socio-histoire des politiques publiques en matiere sociale c~
## 5 "Arthur Devriendt;;LES TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION E~
## 6 "Elmantsr Briak;;Integration forcee de l'afrique subsaharienne dans le proces~
```

```
dim(these_v1)
```

```
## [1] 446871    1
```

```
rm(these_v1)
```

La commande `rm` permet de supprimer la variable `these_v1`

2.2 Jeu de données PhD_v2

2.2.1 Import et préparation des données

```
these <- as_tibble(read_csv("jeux_de_donnees/PhD_v2.csv"))
these$`Date de soutenance` <- dmy(these$`Date de soutenance`)
these$`Publication dans theses.fr` <- dmy(these$`Publication dans theses.fr`)
these$`Mise a jour dans theses.fr` <- dmy(these$`Mise a jour dans theses.fr`)
these$Statut <- as.factor(these$Statut)
these$`Langue de la these` <- as.factor(these$`Langue de la these`)
these$`Accessible en ligne` <- as.factor(these$`Accessible en ligne`)
these$`Identifiant directeur` <- na_if(these$`Identifiant directeur`, "na")
```

2.2.2 Summary.

```
head(these)
```

```
## # A tibble: 6 x 18
##   Auteur      `Identifiant aut~` Titre `Directeur de ~` `Directeur de ~`
##   <chr>      <chr>      <chr> <chr>      <chr>
## 1 Saeed Al marri   <NA>      "Le ~ Philippe Delebe~ Delebecque Phil~
## 2 Andrea Ramazzotti 174423705    "App~ Jean-Claude Gra~ Grandidier Jean~
## 3 OLIVIER BODENREIDER <NA>      "Con~ Francois Kohler  Kohler Francois
## 4 Emmanuel Porte   <NA>      "Soc~ Gilles Pollet   Pollet Gilles
## 5 Arthur Devriendt <NA>      "LES~ Gabriel Dupuy   Dupuy Gabriel
## 6 Elmantsr Briak    <NA>      "Int~ Edmond Jouve    Jouve Edmond
## # ... with 13 more variables: `Identifiant directeur` <chr>,
## #   `Etablissement de soutenance` <chr>, `Identifiant etablisement` <chr>,
## #   Discipline <chr>, Statut <fct>,
## #   `Date de premiere inscription en doctorat` <chr>,
## #   `Date de soutenance` <date>, Year <dbl>, `Langue de la these` <fct>,
## #   `Identifiant de la these` <chr>, `Accessible en ligne` <fct>,
## #   `Publication dans theses.fr` <date>, ...
```

```
## Rows: 447,644
## Columns: 18
## $ Auteur                <chr> "Saeed Al marri", "Andrea R~
## $ `Identifiant auteur`  <chr> NA, "174423705", NA, NA, NA~
## $ Titre                <chr> "Le credit documentaire et ~
## $ `Directeur de these`  <chr> "Philippe Delebecque", "Jea~
## $ `Directeur de these (nom prenom)` <chr> "Delebecque Philippe", "Gra~
## $ `Identifiant directeur` <chr> "29561248", "715,441,511", ~
## $ `Etablissement de soutenance` <chr> "Paris 1", "Chasseneuil-du-~
## $ `Identifiant etablisement` <chr> "27361802", "28024400", NA,~
## $ Discipline            <chr> "Driot prive", "Mecanique d~
## $ Statut                <fct> enCours, enCours, soutenue,~
## $ `Date de premiere inscription en doctorat` <chr> "30-09-11", "01-10-12", NA,~
## $ `Date de soutenance`  <date> NA, NA, 1993-01-01, NA, NA~
## $ Year                  <dbl> NA, NA, 1993, NA, NA, 2008,~
## $ `Langue de la these`  <fct> NA, NA, fr, NA, NA, NA, NA,~
## $ `Identifiant de la these` <chr> "s69480", "s98826", "1993NA~
## $ `Accessible en ligne` <fct> non, non, non, non, non, no~
## $ `Publication dans theses.fr` <date> 2012-01-26, 2013-11-22, 20~
## $ `Mise a jour dans theses.fr` <date> 2012-01-26, 2013-11-22, 20~
```

```
## Auteur      Identifiant auteur  Titre      Directeur de these
## Length:447644  Length:447644  Length:447644  Length:447644
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##
## Directeur de these (nom prenom) Identifiant directeur
## Length:447644      Length:447644
## Class :character    Class :character
## Mode  :character    Mode  :character
##
##
##
##
## Etablissement de soutenance Identifiant etablisement  Discipline
## Length:447644      Length:447644      Length:447644
## Class :character    Class :character    Class :character
## Mode  :character    Mode  :character    Mode  :character
##
##
##
##
## Statut      Date de premiere inscription en doctorat
## enCours : 66329  Length:447644
## soutenue:381315  Class :character
##              Mode  :character
##
##
##
##
## Date de soutenance      Year      Langue de la these
## Min.  :1971-01-01  Min.  :1971  fr  :334404
## 1st Qu.:1994-01-01  1st Qu.:1994  en  :30942
## Median :2004-01-01  Median :2004  enfr :10576
## Mean   :2003-06-08  Mean   :2003  fren :4793
## 3rd Qu.:2012-06-25  3rd Qu.:2012  it   :634
## Max.   :2020-07-07  Max.   :2020  (Other):2530
## NA's   :56746      NA's   :56746  NA's   :63765
## Identifiant de la these Accessible en ligne Publication dans theses.fr
## Length:447644      non:347341      Min.  :2006-04-13
## Class :character    oui:100303      1st Qu.:2013-05-24
## Mode  :character    Median :2013-05-24
##
##              Mean :2014-11-09
##              3rd Qu.:2016-07-11
##              Max.  :2020-07-08
##
##
## Mise a jour dans theses.fr
## Min.  :2010-10-12
## 1st Qu.:2019-04-08
## Median :2020-02-26
## Mean   :2019-06-30
## 3rd Qu.:2020-03-08
## Max.   :2020-07-08
## NA's   :177
```

```
these %>%
  summarise(across(everything(), n_distinct)) %>%
  glimpse()
```

```
## Rows: 1
## Columns: 18
## $ Auteur <int> 430277
## $ `Identifiant auteur` <int> 313775
## $ Titre <int> 446815
## $ `Directeur de these` <int> 159019
## $ `Directeur de these (nom prenom)` <int> 159021
## $ `Identifiant directeur` <int> 98907
## $ `Etablissement de soutenance` <int> 568
## $ `Identifiant etablissement` <int> 573
## $ Discipline <int> 24263
## $ Statut <int> 2
## $ `Date de premiere inscription en doctorat` <int> 4010
## $ `Date de soutenance` <int> 3992
## $ Year <int> 45
## $ `Langue de la these` <int> 206
## $ `Identifiant de la these` <int> 447572
## $ `Accessible en ligne` <int> 2
## $ `Publication dans theses.fr` <int> 2765
## $ `Mise a jour dans theses.fr` <int> 2634
```

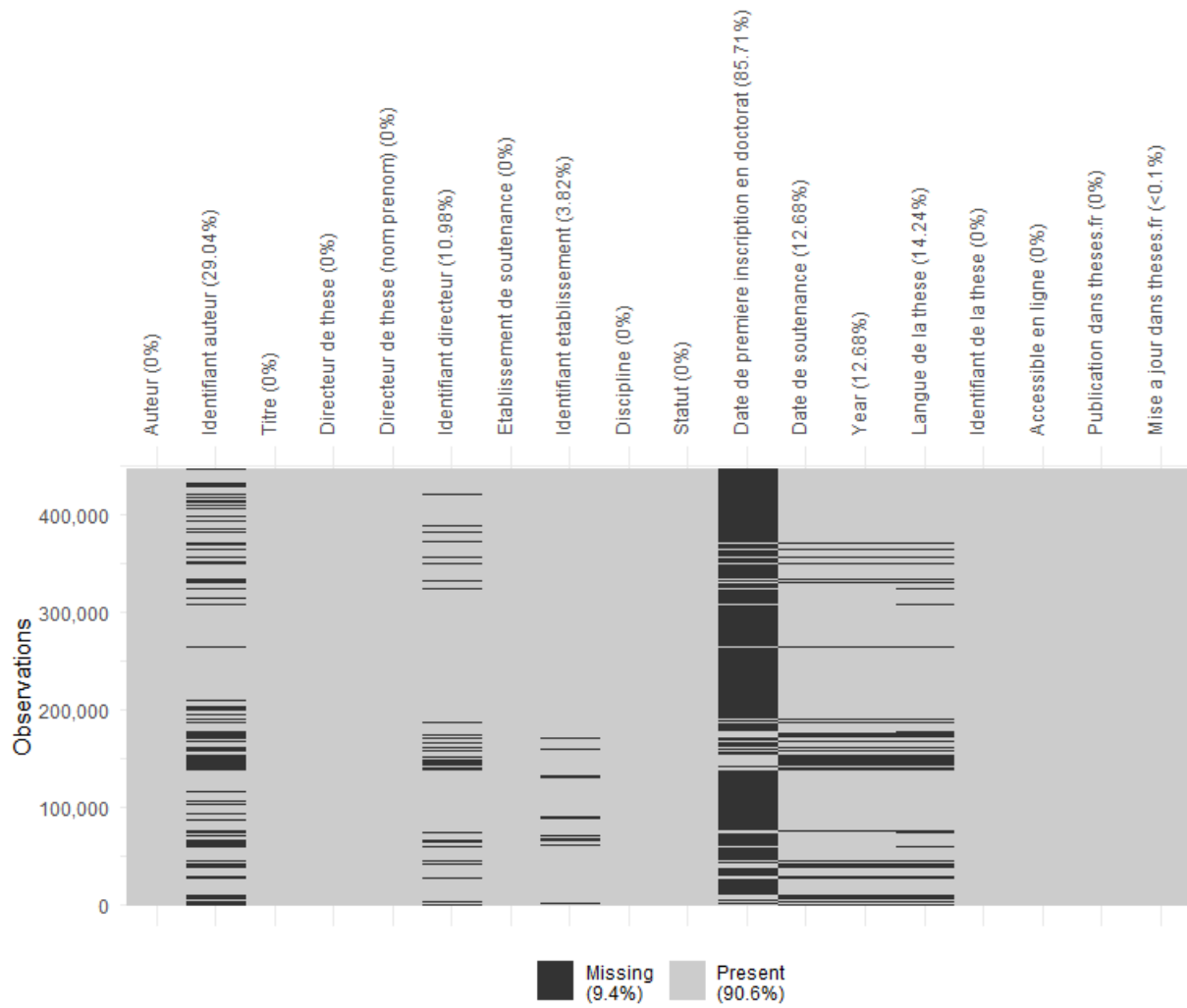
2.3 Visualisation des données manquantes

2.3.1 Préparation des données manquantes

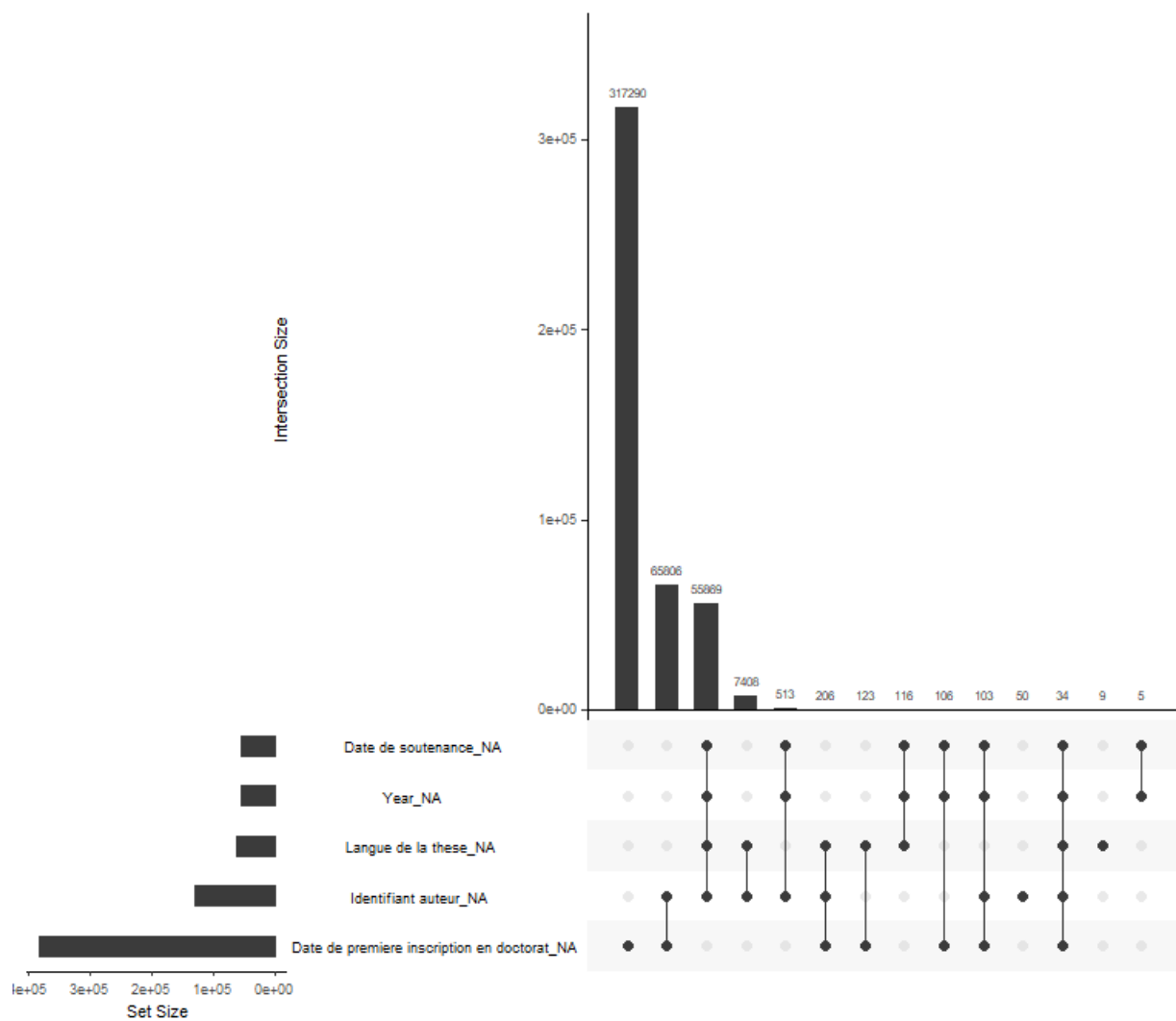
```
these_NA <- these %>% select(Statut, `Date de premiere inscription en doctorat`,
                             `Date de soutenance`, Year, `Langue de la these`)
these_NA_encours <- these_NA %>% filter(Statut == "enCours")
these_NA_soutenue <- these_NA %>% filter(Statut == "soutenue")
```

2.3.2 Visualisation des données manquantes

```
# Visualisation donnees manquantes
vis_miss(these, warn_large_data = FALSE) +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(angle = 90))
```

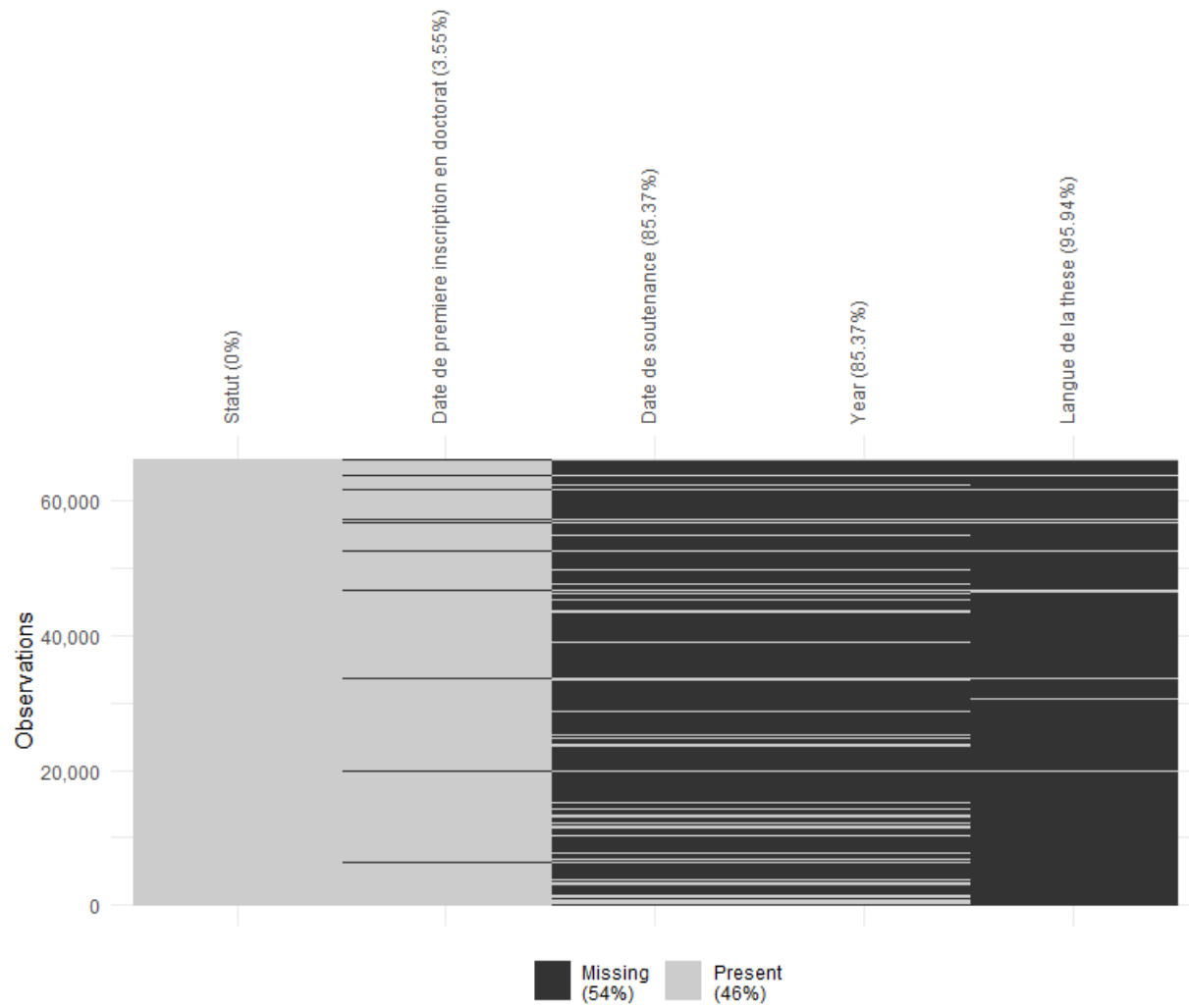


```
gg_miss_upset(these)
```

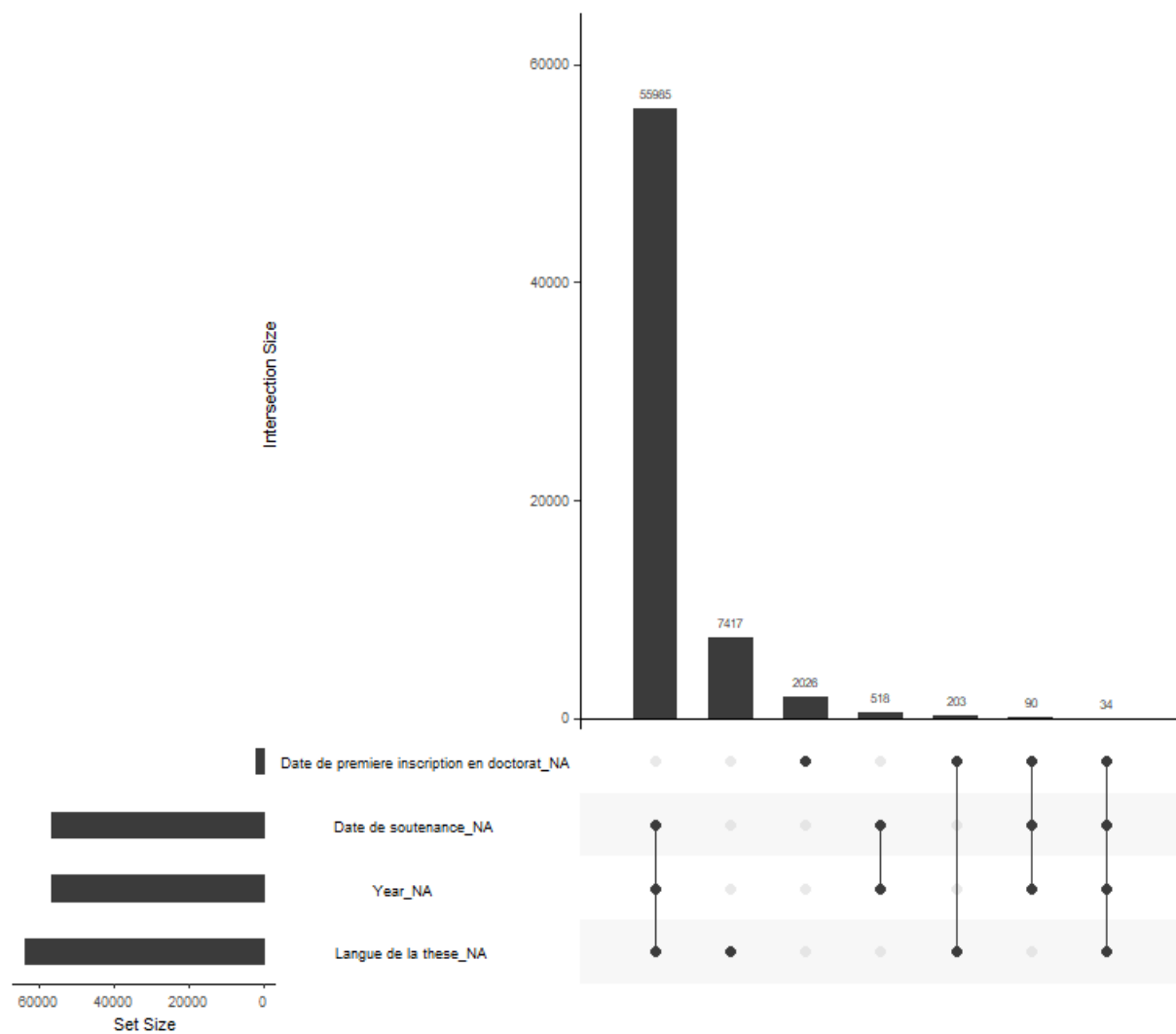


Visualisation statut enCours

```
vis_miss(these_NA_encours, warn_large_data = FALSE) +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(angle = 90))
```

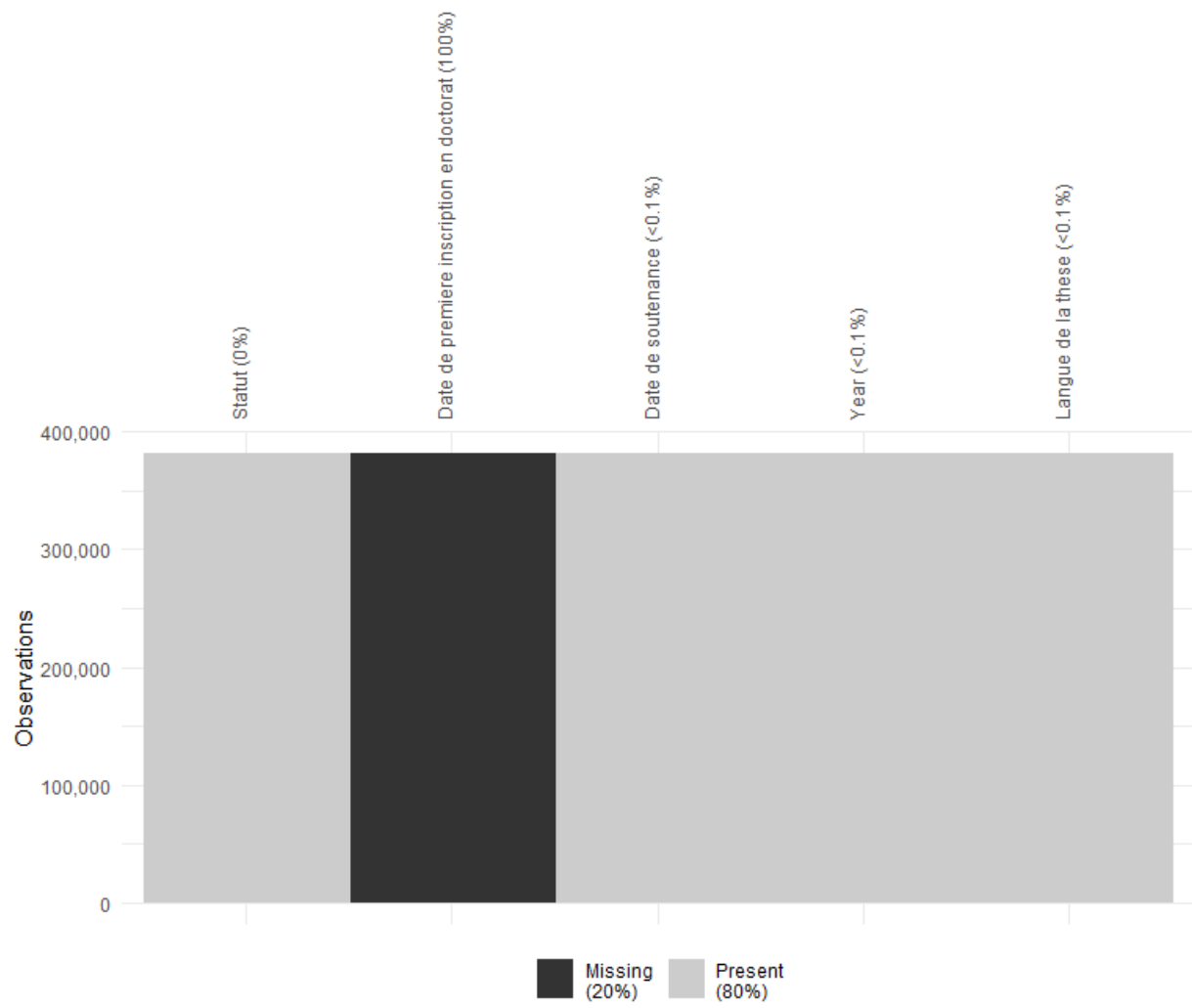


```
gg_miss_upset(these_NA_encours)
```

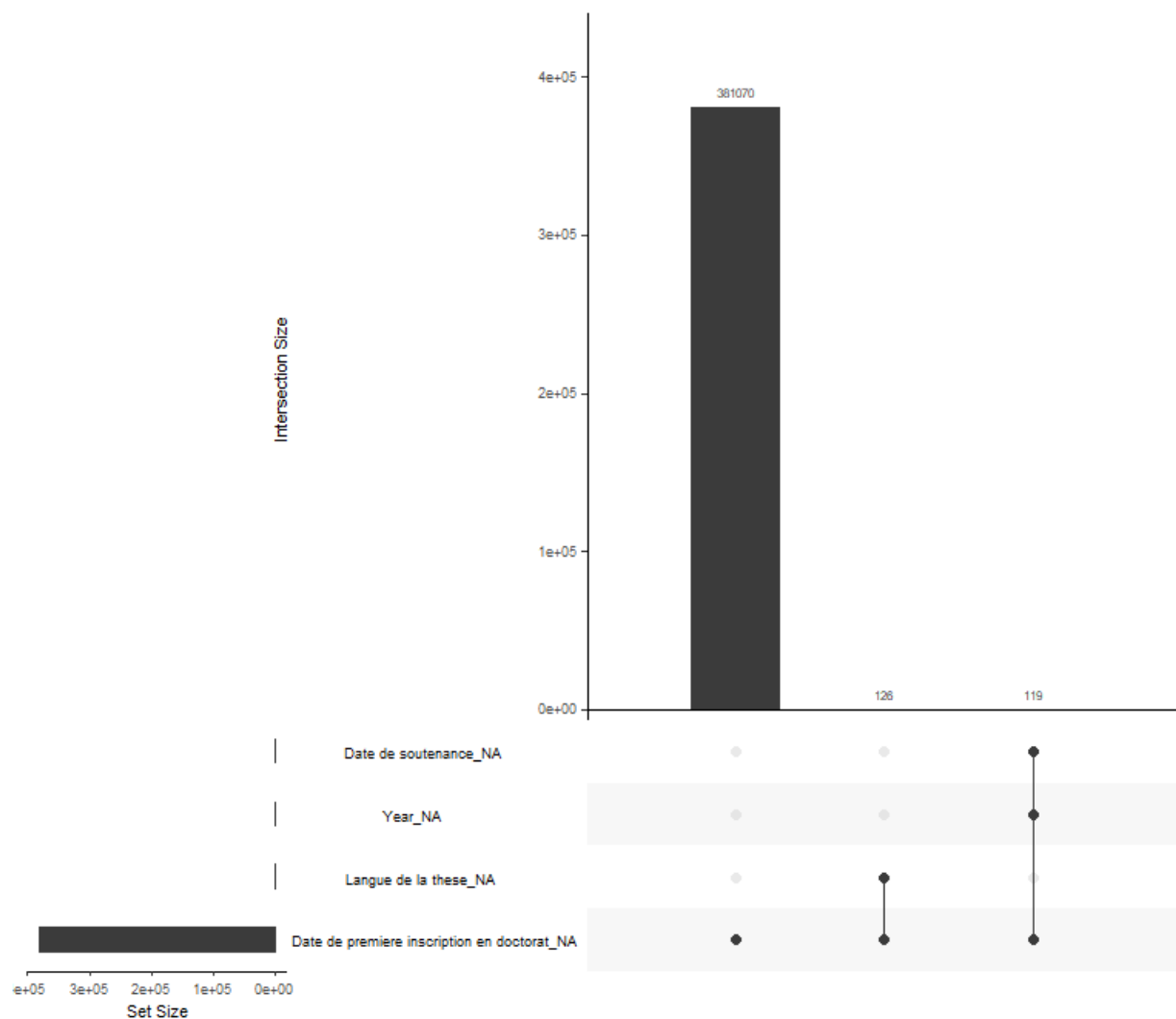



Visualisation statut soutenue

```
vis_miss(these_NA_soutenue, warn_large_data = FALSE) +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(angle = 90))
```



```
gg_miss_upset(these_NA_soutenue)
```



3 Problèmes

3.1 Problèmes soutenances

3.1.1 Préparations des données

```
these_soutenance <- these %>% select(Year, `Date de soutenance`)
these_soutenance <- these_soutenance %>%
  mutate(month = as.factor(month(`Date de soutenance`, label = TRUE)),
         day = as.factor(day(`Date de soutenance`)))
```

Selection des variables cibles et création de nouvelles colonnes `month` et `day` pour plus de lisibilité dans le code suivant.

3.1.2 Préparation des tables intermédiaires.

```
# full_join pour la distribution des pourcentages moyens de soutenance par mois avec errorbar
```

```
# Sans filtrer le 1 Janvier
```

```
these_soutenance_count_year <- these_soutenance %>%  
  filter(Year > 2004 & Year < 2019) %>%  
  count(Year) %>%  
  rename(total_year = n)
```

```
these_soutenance_year_month <- these_soutenance %>%  
  filter(Year > 2004 & Year < 2019) %>%  
  count(Year, month) %>%  
  rename(total_month = n)
```

```
these_soutenance_full <- full_join(these_soutenance_count_year,  
                                   these_soutenance_year_month,  
                                   by = "Year") %>%  
  mutate(freq = total_month / total_year) %>%  
  drop_na()
```

```
# Avec filtre sur le 1 Janvier
```

```
these_soutenance_count_year_no_first <- these_soutenance %>%  
  filter(day != "1" & Year > 2004 & Year < 2019) %>%  
  count(Year) %>%  
  rename(total_year = n)
```

```
these_soutenance_year_month_no_first <- these_soutenance %>%  
  filter(day != "1" & Year > 2004 & Year < 2019) %>%  
  count(Year, month) %>%  
  rename(total_month = n)
```

```
these_soutenance_full_no_first <- full_join(these_soutenance_count_year_no_first,  
                                             these_soutenance_year_month_no_first,  
                                             by = "Year") %>%  
  mutate(freq = total_month / total_year) %>%  
  drop_na()
```

```
# full_join pour l'evolution des frequences des defenses de theses par annees au premier janvier
```

```
these_soutenance_count_year <- these_soutenance %>%  
  count(Year) %>%  
  rename(total_year = n)
```

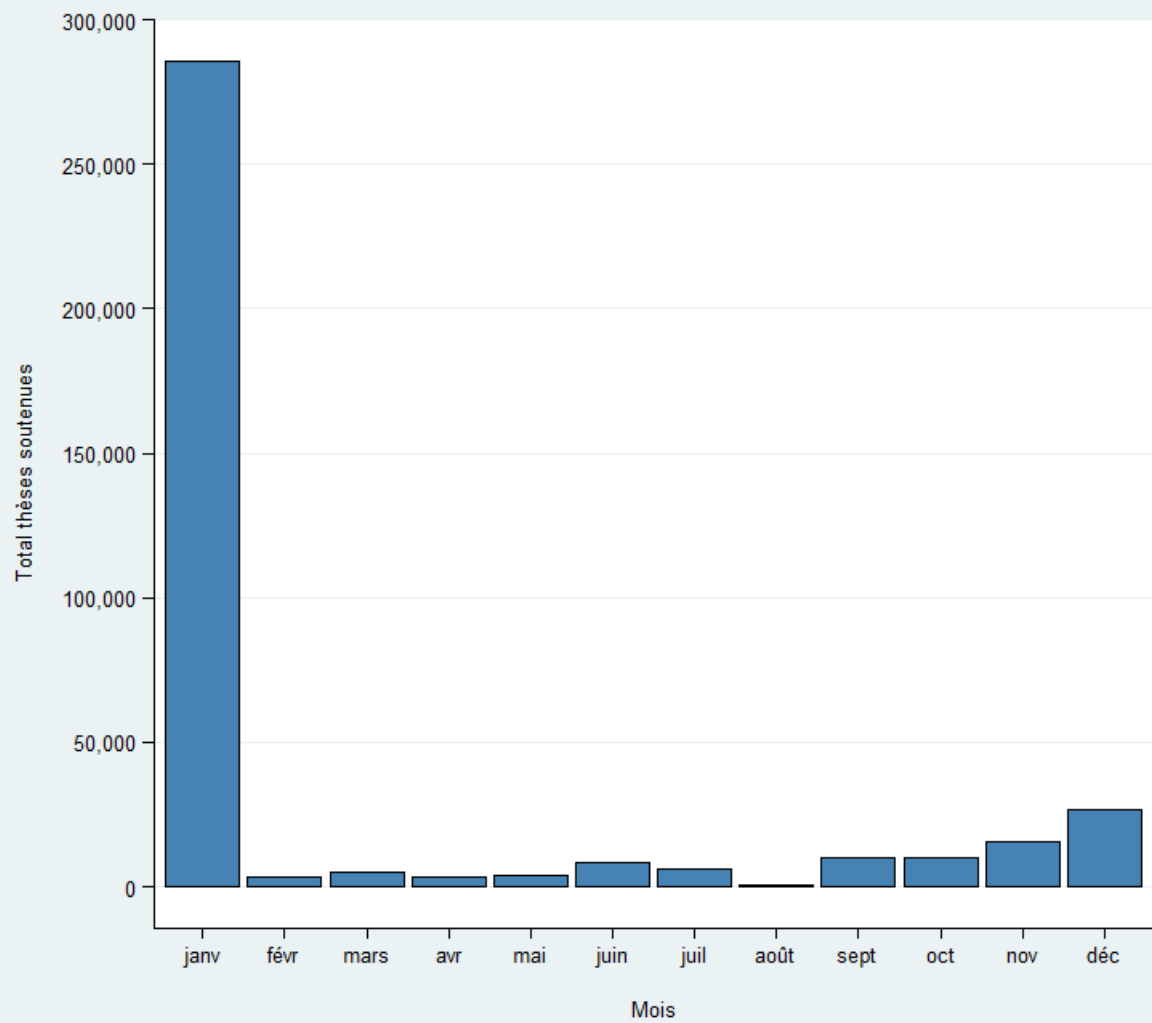
```
these_soutenance_count_janv <- these_soutenance %>%  
  filter(month == "janv" & day == "1") %>%  
  group_by(Year) %>%  
  count(Year) %>%  
  rename(total_janv = n)
```

```
these_soutenance_year_janv <- full_join(these_soutenance_count_year,  
                                         these_soutenance_count_janv, by = "Year") %>%  
  mutate(freq = total_janv / total_year)
```

3.1.3 Visualisations

```
# Distribution du total these par mois
```

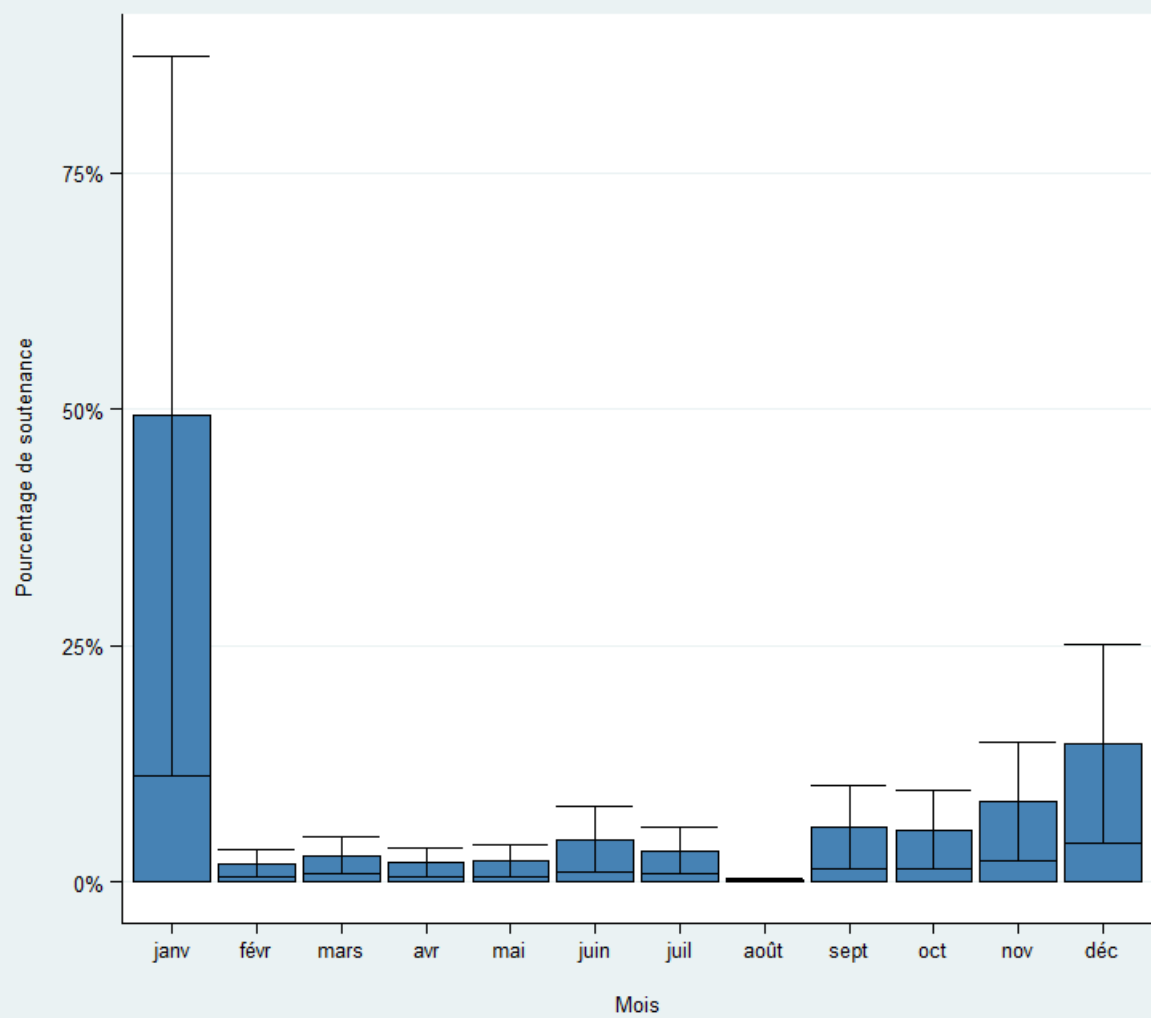
```
these_soutenance %>%  
  filter(Year > 1983 & Year < 2019) %>%  
  count(month) %>%  
  ggplot(aes(month, n)) +  
  geom_col(fill = "steelblue", color = "black") +  
  scale_y_continuous(labels = comma,  
                     breaks = seq(0, 300000, 50000)) +  
  labs(x = "\nMois",  
       y = "Total thèses soutenues\n") +  
  theme_stata() +  
  theme(axis.text.y = element_text(angle = 0))
```



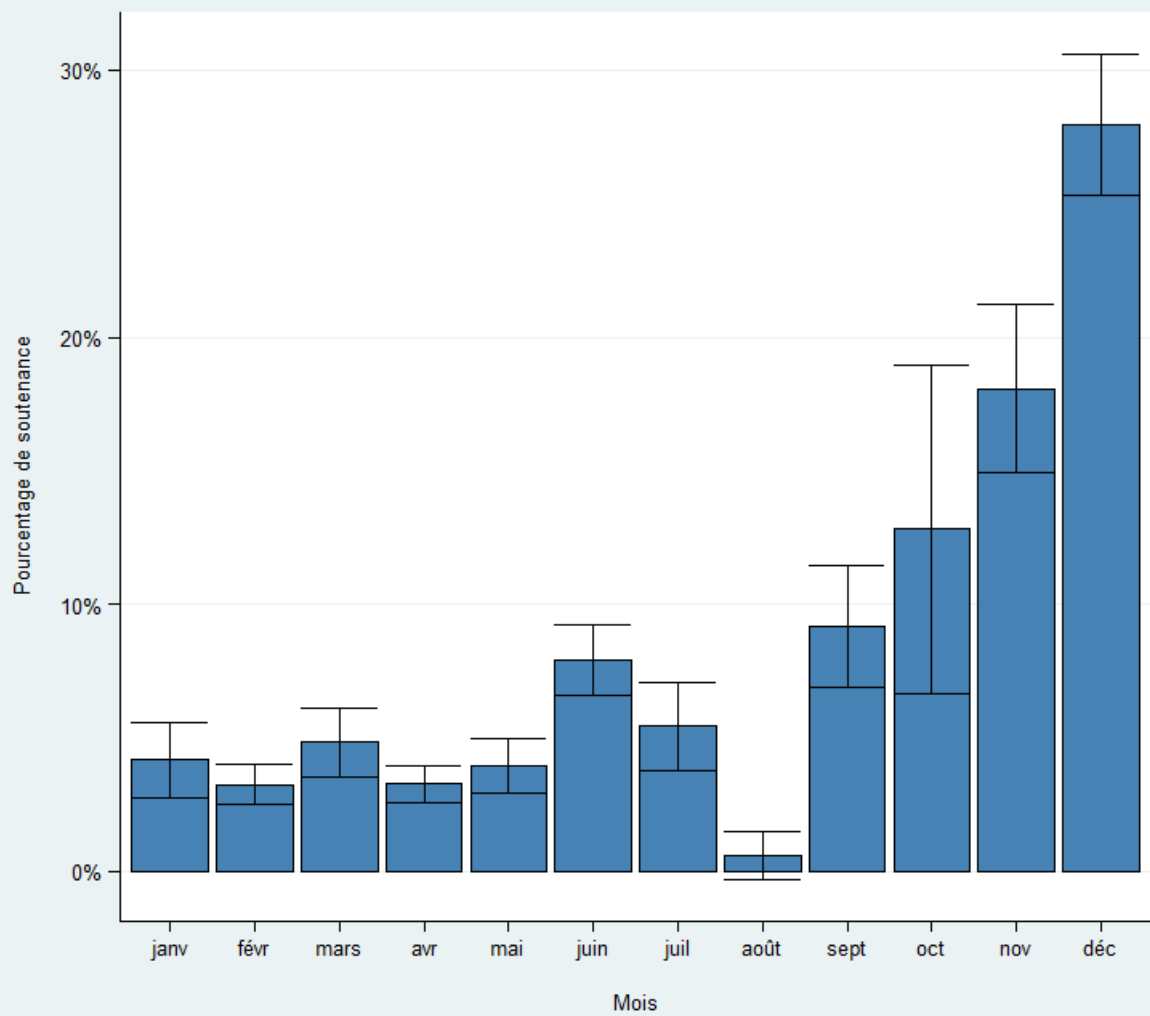
```
# facet_wrap month~year
these_soutenance %>%
  filter(Year > 2004 & Year < 2019) %>%
  count(Year, month) %>%
  ggplot(aes(month, n)) +
  geom_col(fill = "steelblue", color = "black") +
  facet_wrap( ~ Year) +
  scale_x_discrete(guide = guide_axis(angle = 90)) +
  labs(x = "\nMois",
       y = "Total thèses soutenues\n") +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 0))
```



```
# Distribution des pourcentages moyens de soutenance par mois avec errorbar
# Avec 1 Janvier
these_soutenance_full %>%
  ddply(~month, summarise, mean = mean(freq, na.rm = TRUE), sd = sd(freq, na.rm = TRUE)) %>%
  ggplot(aes(month, mean)) +
  geom_col(fill = "steelblue", color = "black") +
  geom_errorbar(aes(ymin = mean - sd, ymax = mean + sd)) +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "\nMois",
       y = "Pourcentage de soutenance\n") +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 0))
```



```
# Sans 1 Janvier
these_soutenance_full_no_first %>%
  ddply(~month, summarise, mean = mean(freq, na.rm = TRUE), sd = sd(freq, na.rm = TRUE)) %>%
  ggplot(aes(month, mean)) +
  geom_col(fill = "steelblue", color = "black") +
  geom_errorbar(aes(ymin = mean - sd, ymax = mean + sd)) +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "\nMois",
       y = "Pourcentage de soutenance\n") +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 0))
```



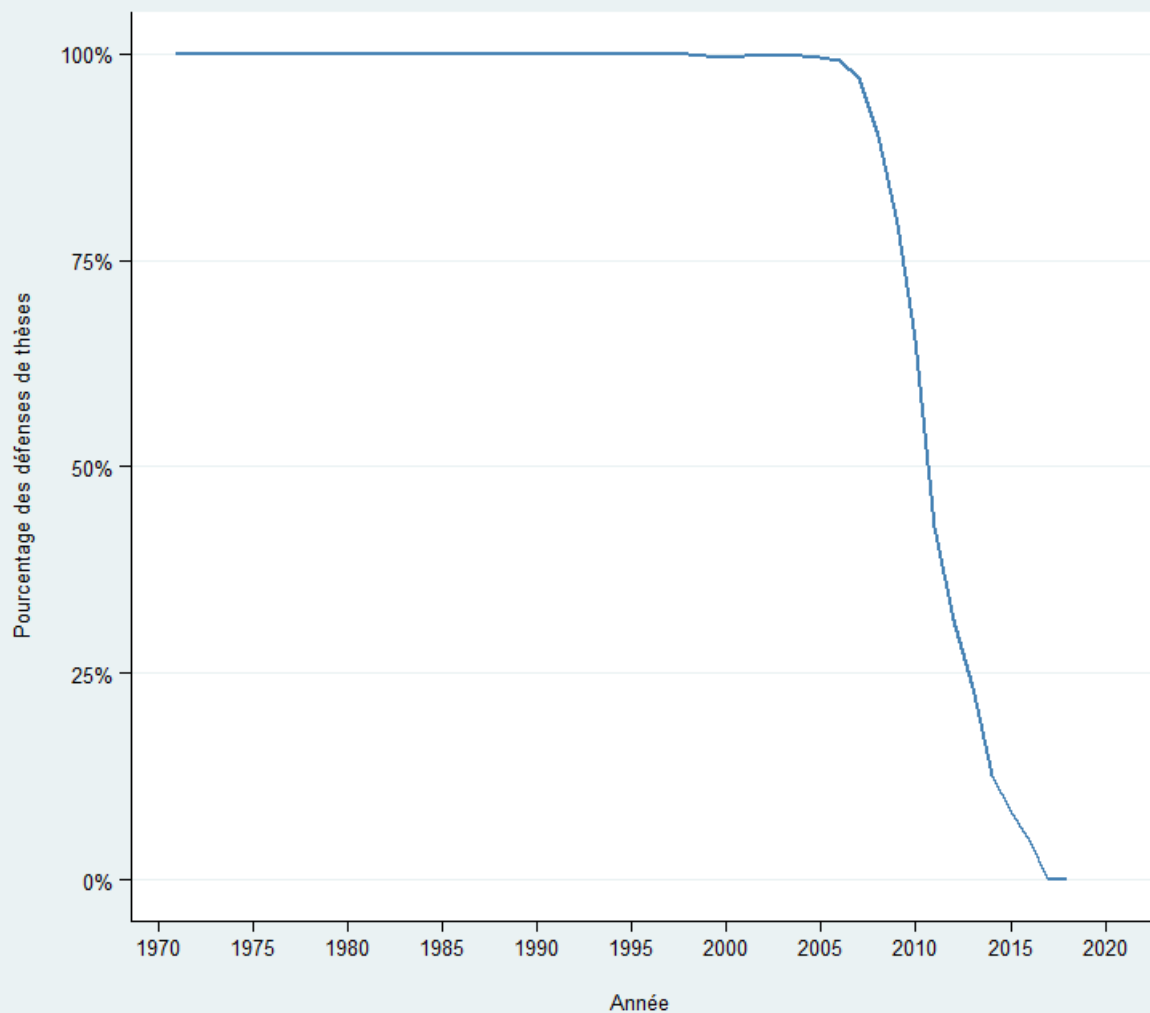
Evolution des frequences des defenses de theses par annees au premier janvier

`these_soutenance_year_janv %>%`

```

ggplot(aes(Year, freq)) +
  geom_line(color = "steelblue", size = 1) +
  scale_x_continuous(breaks = seq(1970, 2020, 5)) +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "\nAnnée",
       y = "Pourcentage des défenses de thèses\n") +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 0))

```

3.2 Problème homonyme Cecile Martin

3.2.1 Import et préparation des données

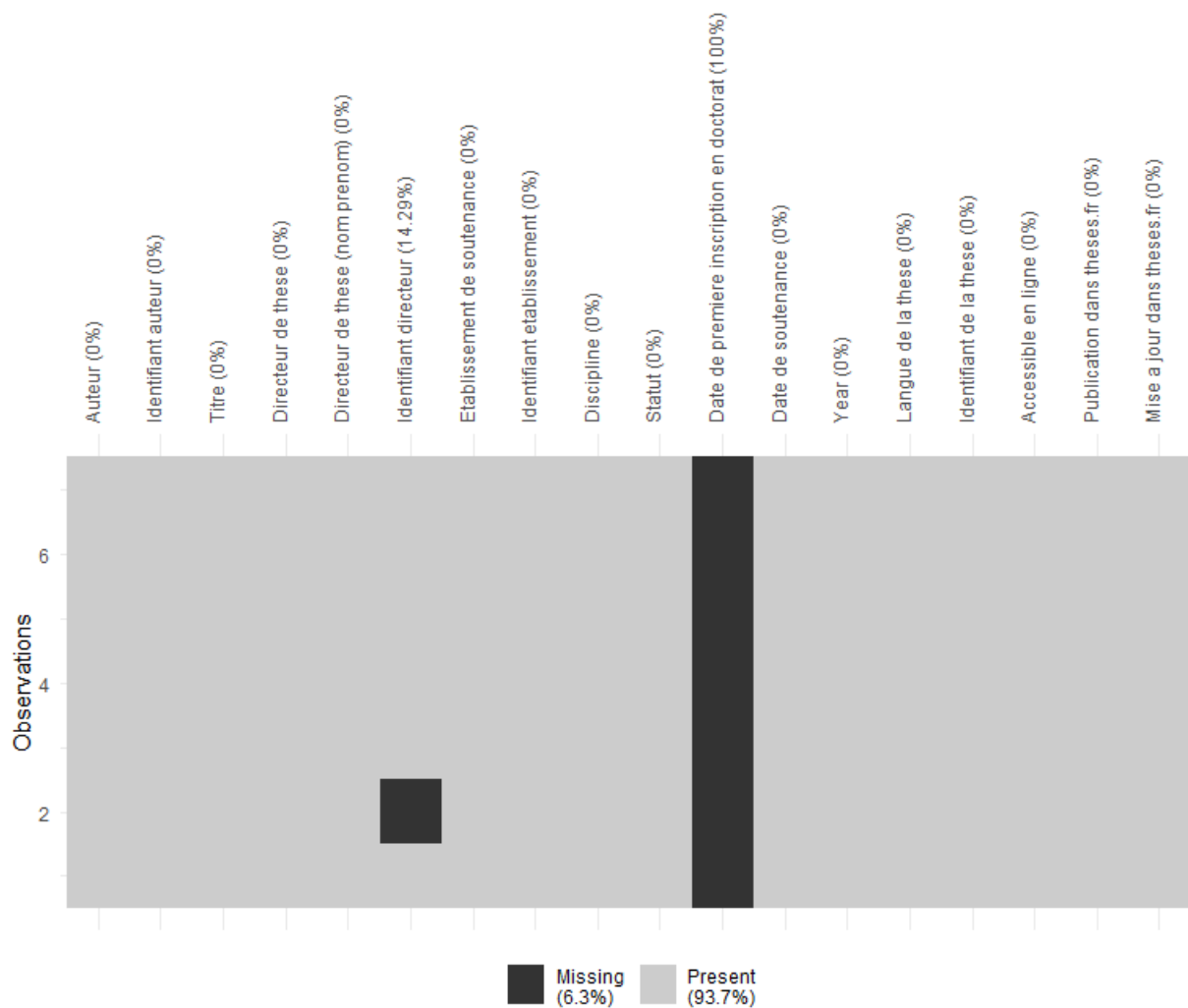
```
these_cecile_martin <- these %>%
  filter(str_detect(Auteur, "Cecile Martin")) %>%
  slice(-c(4,6, 8, 9, 12))
```

```
glimpse(these_cecile_martin %>%
  summarise(across(everything(), n_distinct)))
```

```
## Rows: 1
## Columns: 18
## $ Auteur <int> 1
## $ `Identifiant auteur` <int> 4
## $ Titre <int> 7
## $ `Directeur de these` <int> 7
## $ `Directeur de these (nom prenom)` <int> 7
## $ `Identifiant directeur` <int> 7
## $ `Etablissement de soutenance` <int> 7
## $ `Identifiant etablissement` <int> 7
## $ Discipline <int> 7
## $ Statut <int> 1
## $ `Date de premiere inscription en doctorat` <int> 1
## $ `Date de soutenance` <int> 7
## $ Year <int> 7
## $ `Langue de la these` <int> 2
## $ `Identifiant de la these` <int> 7
## $ `Accessible en ligne` <int> 2
## $ `Publication dans theses.fr` <int> 3
## $ `Mise a jour dans theses.fr` <int> 5
```

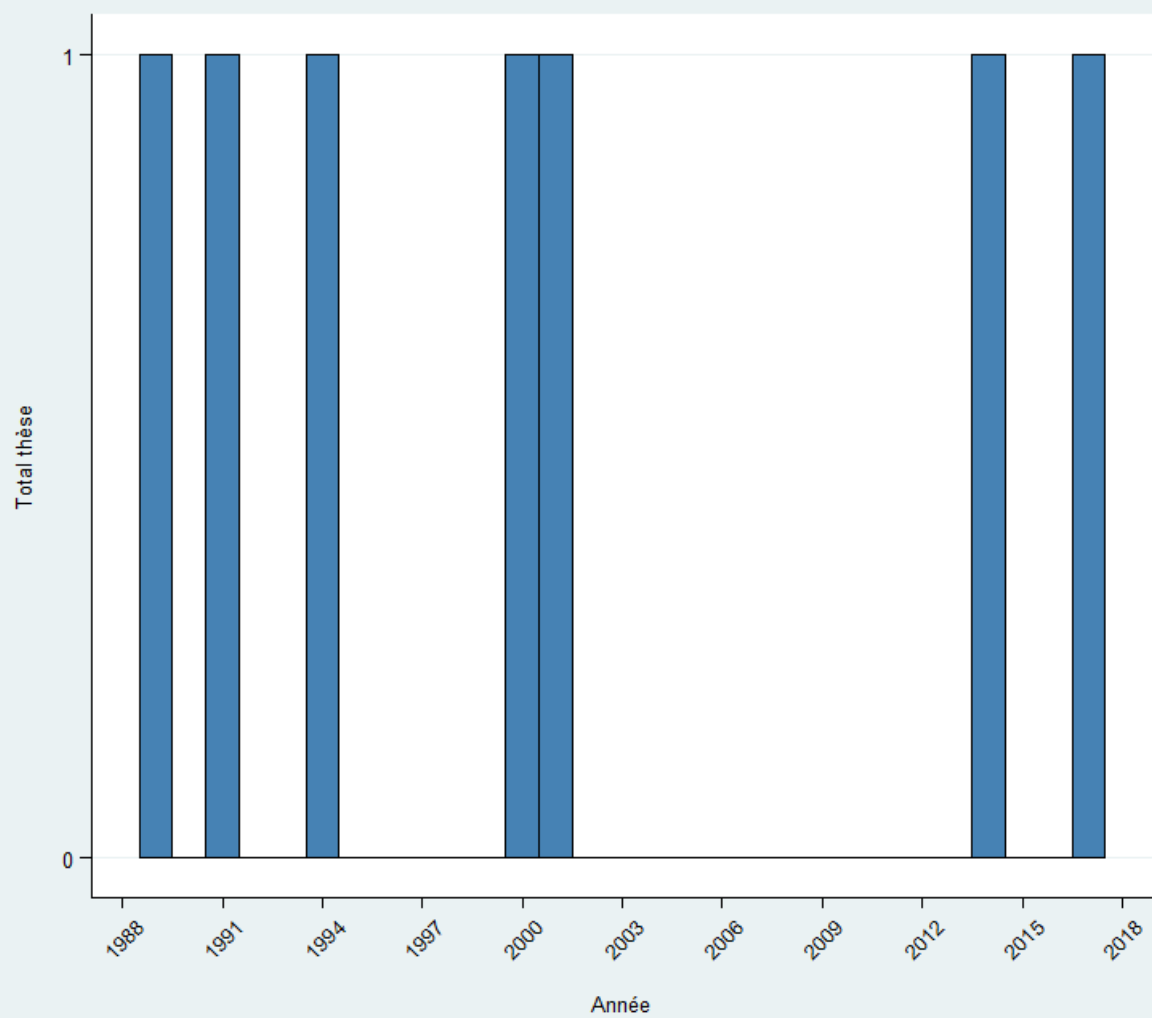
3.2.2 Visualisation des données manquantes

```
these_cecile_martin %>%
  vis_miss() +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(angle = 90))
```

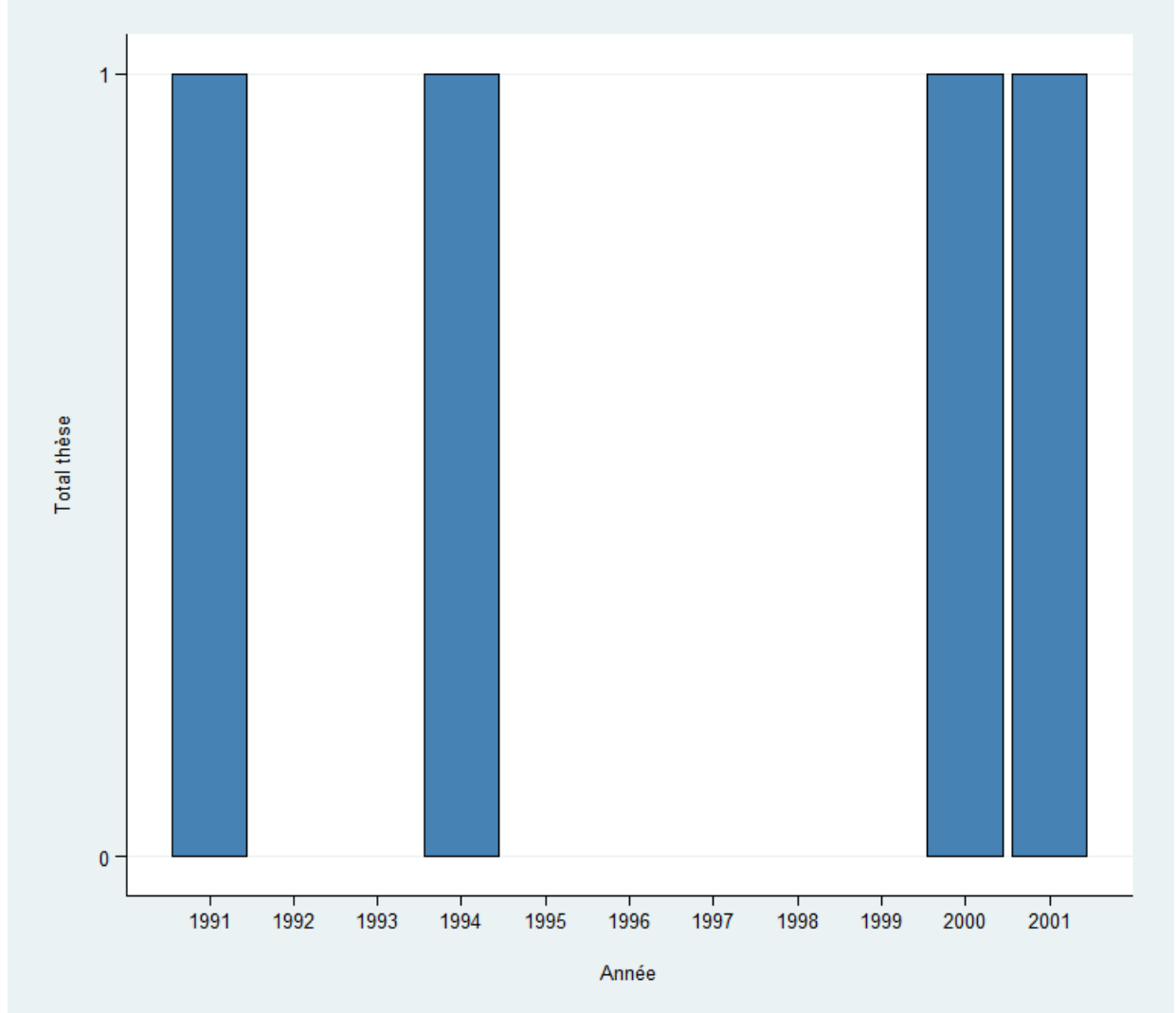


3.2.3 Exploration des données

```
these_cecile_martin %>%
  ggplot(aes(x = Year)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black") +
  scale_x_continuous(breaks = seq(1985, 2020, 3)) +
  scale_y_continuous(breaks = c(0, 1)) +
  labs(x = "\nAnnée",
       y = "Total thèse\n") +
  theme_stata() +
  theme(axis.text.x = element_text(angle = 45,
                                    vjust = 0.5),
        axis.text.y = element_text(angle = 0))
```



```
# Filtrage sur l'Identifiant auteur "81323557"
these_cecile_martin %>%
  filter(`Identifiant auteur` == "81323557") %>%
  ggplot(aes(x = Year)) +
  geom_bar(binwidth = 1, fill = "steelblue", color = "black") +
  scale_x_continuous(breaks = seq(1990, 2002, 1)) +
  scale_y_continuous(breaks = c(0, 1)) +
  labs(x = "\nAnnée",
       y = "Total thèse\n") +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 0))
```



4 Outliers

4.1 Outliers director

4.1.1 Import et préparation des données

```
these_director <- these[!grepl(",", these$`Directeur de these (nom prenom)`), ]
these_director <- these_director[!grepl("@", these_director$`Directeur de these (nom prenom)`), ]
```

```
these_director <- these_director %>%
  filter(Year > 1983 & Year < 2019) %>%
  select(`Directeur de these (nom prenom)`, `Identifiant directeur`) %>%
  group_by(`Directeur de these (nom prenom)`) %>%
  mutate(total_these_diriger = n())
```

```
dim(these_director)
```

```
## [1] 308587    3
```

```
n_distinct(these_director$`Directeur de these (nom prenom)`)
```

```
## [1] 66148
```

```
n_distinct(these_director$`Identifiant directeur`)
```

```
## [1] 58680
```

4.1.2 Méthodes de différent moyen de calcul des outliers

```
# Mean and Standard deviation (SD)
## Tmin, Tmax = mean(+-(k*sd), k = 3
Tmin_msd <- mean(these_director$total_these_diriger) - ( 3 * sd(these_director$total_these_diriger))
Tmax_msd <- mean(these_director$total_these_diriger) + ( 3 * sd(these_director$total_these_diriger))
msd <- which(these_director$total_these_diriger < Tmin_msd |
              these_director$total_these_diriger > Tmax_msd)
msd_outliers <- these_director[msd,]
min(msd_outliers$total_these_diriger)
```

```
## [1] 131
```

```
max(msd_outliers$total_these_diriger)
```

```
## [1] 711
```

```
# Median and Median Absolute Deviation (MAD)
## MAD = b * median(|xi - median(x)|), b = 1.4826 for normal distribution
## Tmin, Tmax = median(+-(k*MAD), k = 3
med <- median(these_director$total_these_diriger)
abs_med <- abs(these_director$total_these_diriger - med)
mad <- 1.4826 * median(abs_med)
Tmin_mad <- med - ( 3 * mad)
Tmax_mad <- med + ( 3 * mad)
mad <- which(these_director$total_these_diriger < Tmin_mad |
              these_director$total_these_diriger > Tmax_mad)
mad_outliers <- these_director[mad,]
min(mad_outliers$total_these_diriger)
```

```
## [1] 36
```

```
max(mad_outliers$total_these_diriger)
```

```
## [1] 711
```

```
# Interquartile Range (IQR)
## Tmin = Q1 - (c * IQR) c = 1.5(mild)
## Tmax = Q3 + (c * IQR) c = 1.5(mild)
summary(these_director)
```

```
## Directeur de these (nom prenom) Identifiant directeur total_these_diriger
## Length:308587      Length:308587      Min.   : 1.00
## Class :character   Class :character  1st Qu.: 4.00
## Mode  :character   Mode :character   Median : 9.00
##                      Mean    :16.21
##                      3rd Qu.:18.00
##                      Max.    :711.00
```

```
IQR(these_director$total_these_diriger)
```

```
## [1] 14
```

```
Tmin_iqr_mild <- 4 - (1.5 * 14)
Tmax_iqr_mild <- 18 + (1.5 * 14)
iqr_mild <- which(these_director$total_these_diriger < Tmin_iqr_mild |
                 these_director$total_these_diriger > Tmax_iqr_mild)
iqr_mild_outliers <- these_director[iqr_mild,]
min(iqr_mild_outliers$total_these_diriger)
```

```
## [1] 40
```

```
max(iqr_mild_outliers$total_these_diriger)
```

```
## [1] 711
```

```
## Tmin = Q1 - (c * IQR) c = 3(extreme)
## Tmax = Q3 + (c * IQR) c = 3(extreme)
Tmin_iqr_ext <- 4 - (3 * 14)
Tmax_iqr_ext <- 18 + (3 * 14)
iqr_ext <- which(these_director$total_these_diriger < Tmin_iqr_ext |
                 these_director$total_these_diriger > Tmax_iqr_ext)
iqr_ext_outliers <- these_director[iqr_ext,]
min(iqr_ext_outliers$total_these_diriger)
```

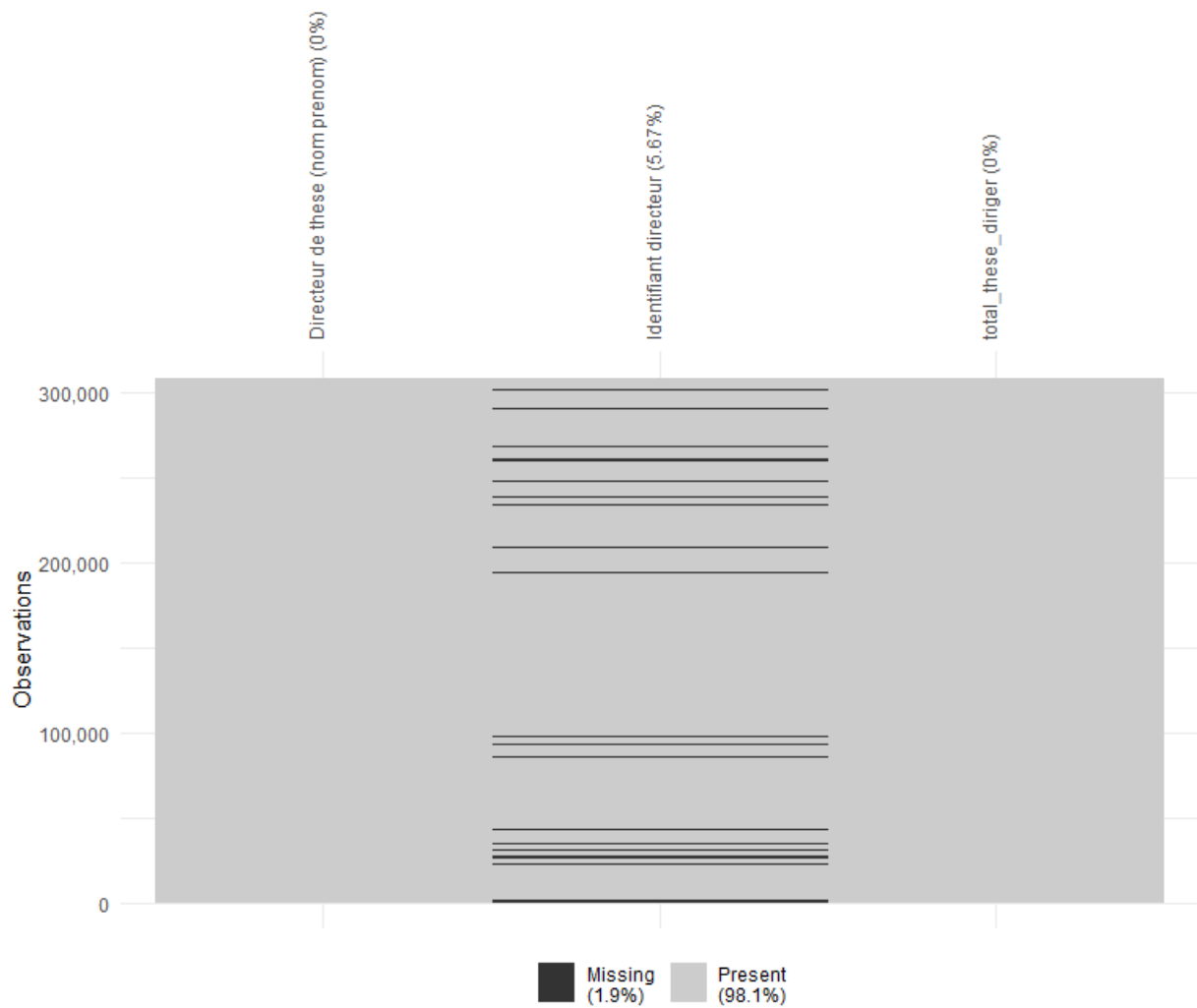
```
## [1] 61
```

```
max(iqr_ext_outliers$total_these_diriger)
```

```
## [1] 711
```

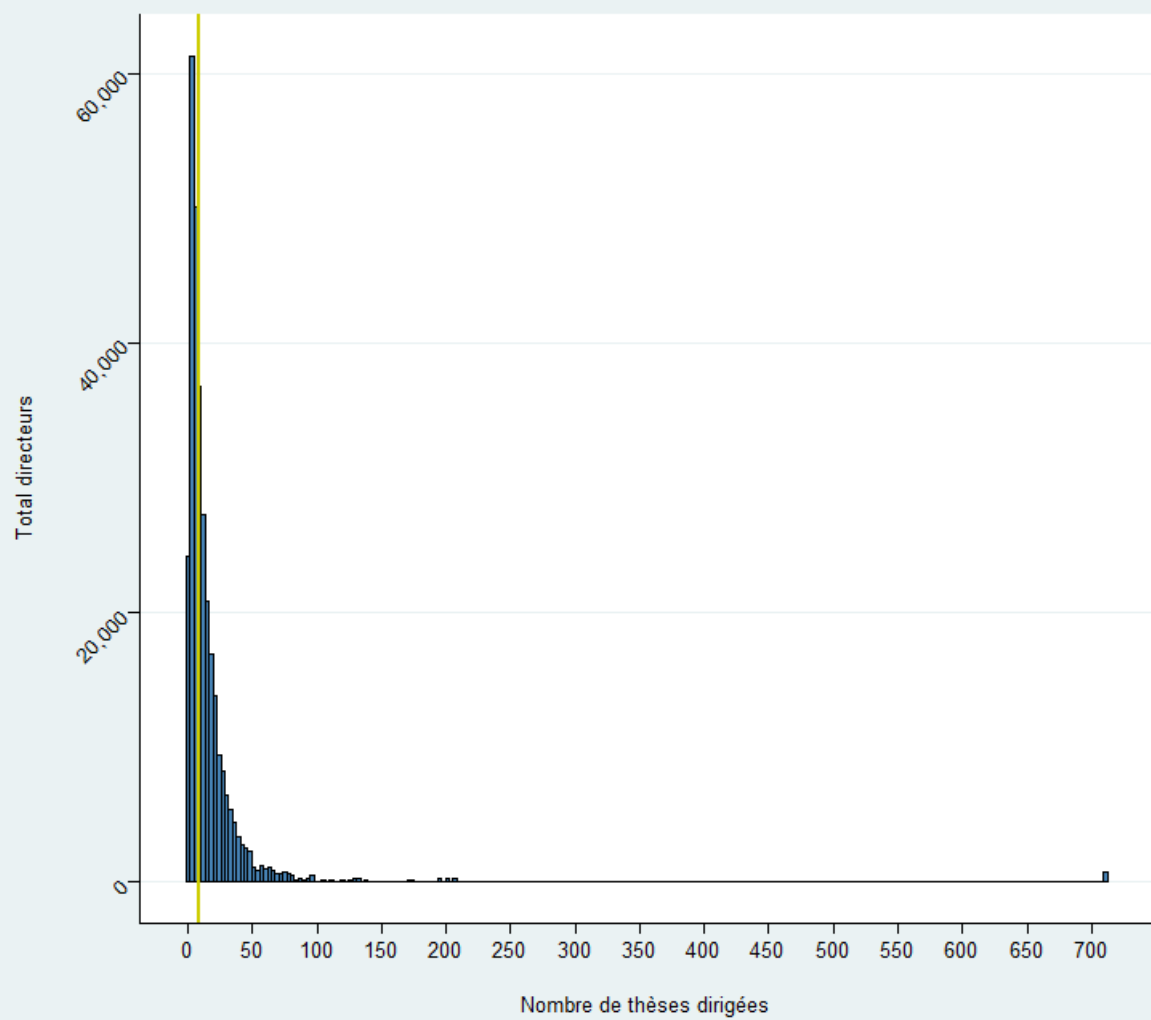
4.1.3 Visualisation des données manquantes

```
these_director %>% vis_miss(warn_large_data = FALSE) +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(angle = 90))
```

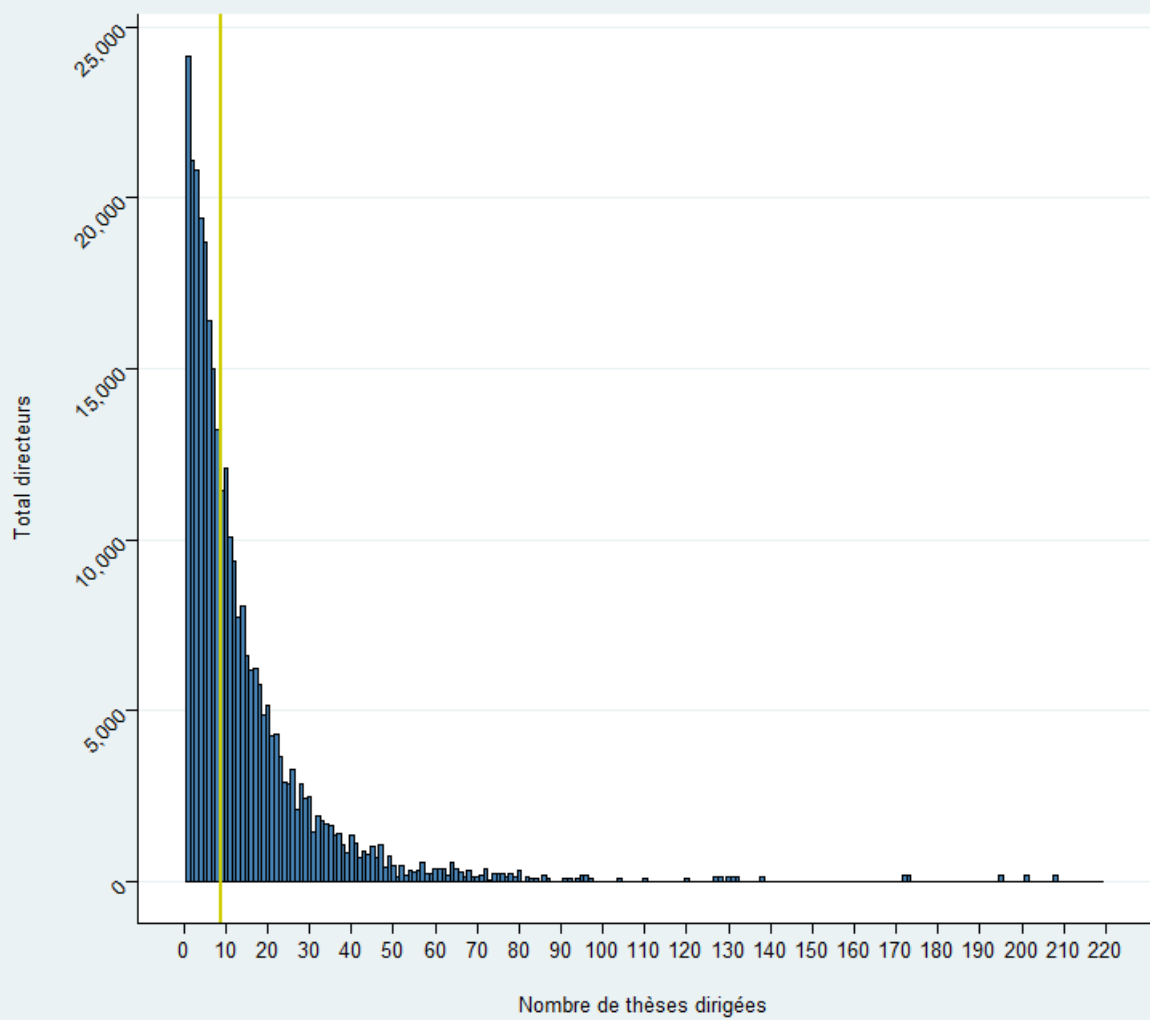


4.1.4 Recherches des outliers

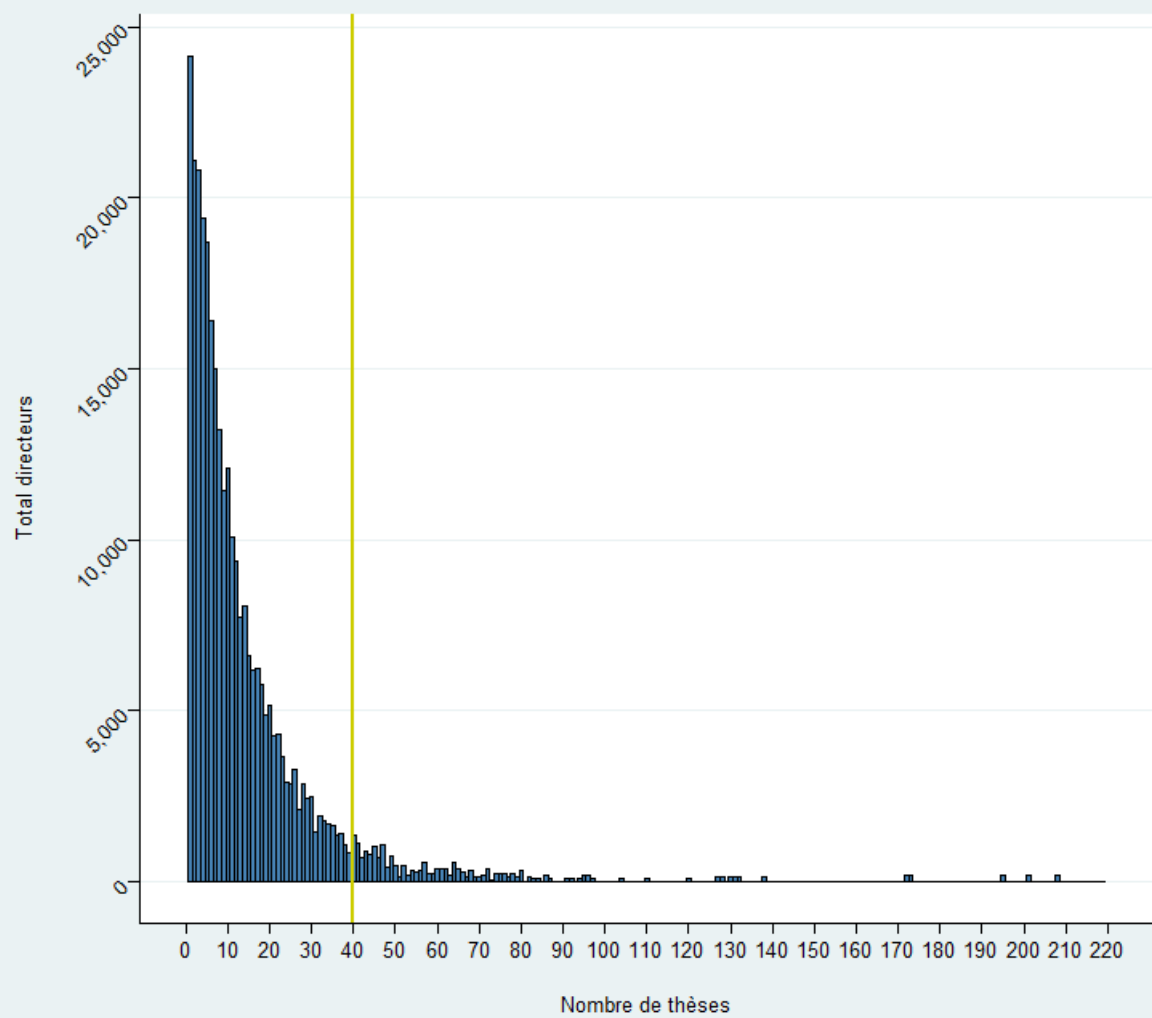
```
these_director %>%
  ggplot(aes(total_these_diriger)) +
  geom_histogram(binwidth = 3, fill = "steelblue", color = "Black") +
  geom_vline(aes(xintercept = median(total_these_diriger)), color = "yellow3", size = 1) +
  scale_y_continuous(labels = comma) +
  labs(x = "\nNombre de thèses dirigées",
       y = "Total directeurs\n") +
  scale_x_continuous(breaks = seq(0, 800, 50)) +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 45))
```



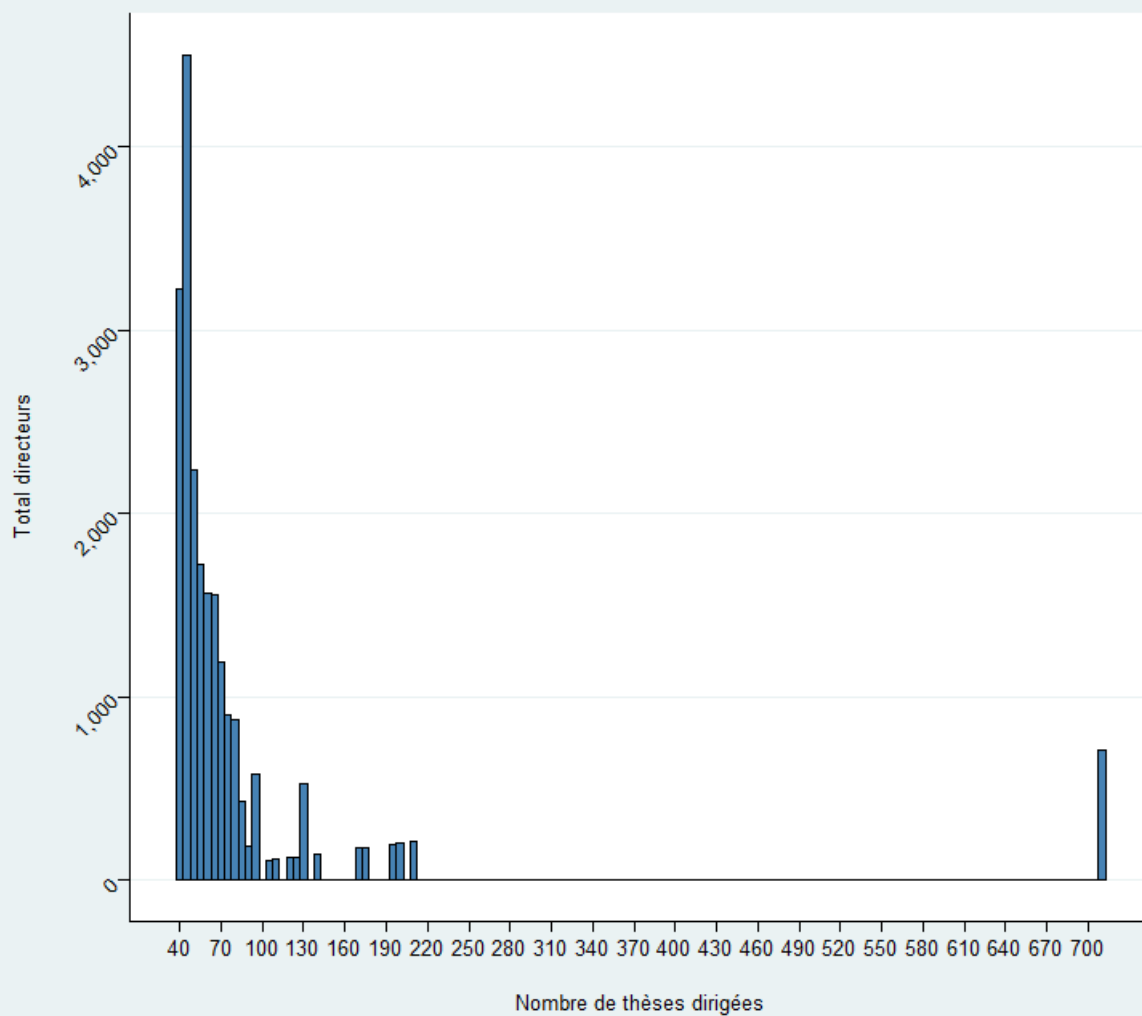
```
these_director %>%
  ggplot(aes(total_these_diriger)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "Black") +
  geom_vline(aes(xintercept = median(total_these_diriger)), color = "yellow3", size = 1) +
  scale_x_continuous(limits = c(0, 220), breaks = seq(0, 220, 10)) +
  scale_y_continuous(labels = comma) +
  labs(x = "\nNombre de thèses dirigées",
       y = "Total directeurs\n") +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 45))
```

```
these_director %>%
  ggplot(aes(total_these_diriger)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "Black") +
  geom_vline(aes(xintercept = min(iqr_mild_outliers$total_these_diriger)), color = "yellow3", size = 1) +
  scale_x_continuous(limits = c(0, 220), breaks = seq(0, 220, 10)) +
  scale_y_continuous(labels = comma) +
  labs(x = "\nNombre de thèses",
       y = "Total directeurs\n") +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 45))
```



```
these_director %>%
  filter(total_these_diriger >= 40) %>%
  ggplot(aes(total_these_diriger)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "Black") +
  scale_x_continuous(breaks = seq(40, 720, 30)) +
  scale_y_continuous(labels = comma) +
  labs(x = "\nNombre de thèses dirigées",
       y = "Total directeurs\n") +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 45))
```



4.1.5 Analyse des outliers

```
# Total theses >= 40 & <= 140
these_director_outliers <- these_director %>% filter(total_these_diriger >= 40 &
                                                    total_these_diriger <= 140)
dim(these_director_outliers)
```

```
## [1] 20068    3
```

```
n_distinct(these_director_outliers$`Directeur de these (nom prenom)`)
```

```
## [1] 367
```

```
n_distinct(these_director_outliers$`Identifiant directeur`)
```

```
## [1] 447
```

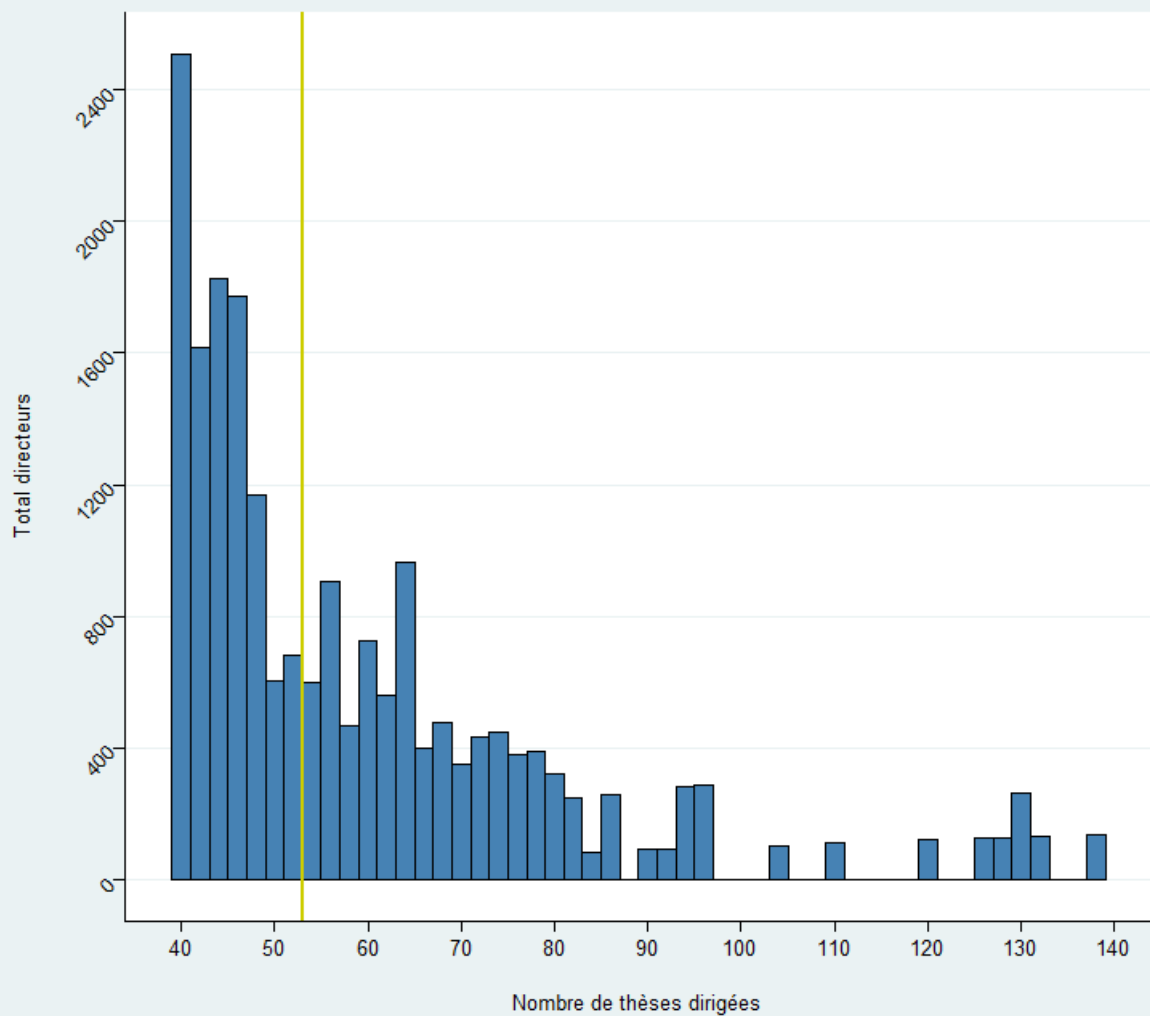
```
glimpse(these_director_outliers)
```

```
## Rows: 20,068
## Columns: 3
## Groups: Directeur de these (nom prenom) [367]
## $ `Directeur de these (nom prenom)` <chr> "Jouve Edmond", "Dekeuwer-Defossez F~
## $ `Identifiant directeur`      <chr> "26941848", "26818094", "26941848", ~
## $ total_these_diriger         <int> 46, 44, 46, 41, 59, 44, 75, 75, 43, ~
```

```
vis_miss(these_director_outliers)
```



```
ggplot(these_director_outliers, aes(total_these_diriger)) +
  geom_histogram(binwidth = 2, fill = "steelblue", color = "Black") +
  geom_vline(aes(xintercept = median(total_these_diriger)), color = "yellow3", size = 1) +
  labs(x = "\nNombre de thèses dirigées",
       y = "Total directeurs\n") +
  scale_x_continuous(breaks = seq(40, 140, 10)) +
  scale_y_continuous(breaks = seq(0, 2600, 400)) +
  theme_stata() +
  theme(axis.text.y = element_text(angle = 45))
```



```
# total theses >=140 & <= 240
these_director_outliers_middle <- these_director %>% filter(total_these_diriger >= 140 &
  total_these_diriger <= 240)

dim(these_director_outliers_middle)
```

```
## [1] 949 3
```

```
n_distinct(these_director_outliers_middle$`Directeur de these (nom prenom)`)
```

```
## [1] 5
```

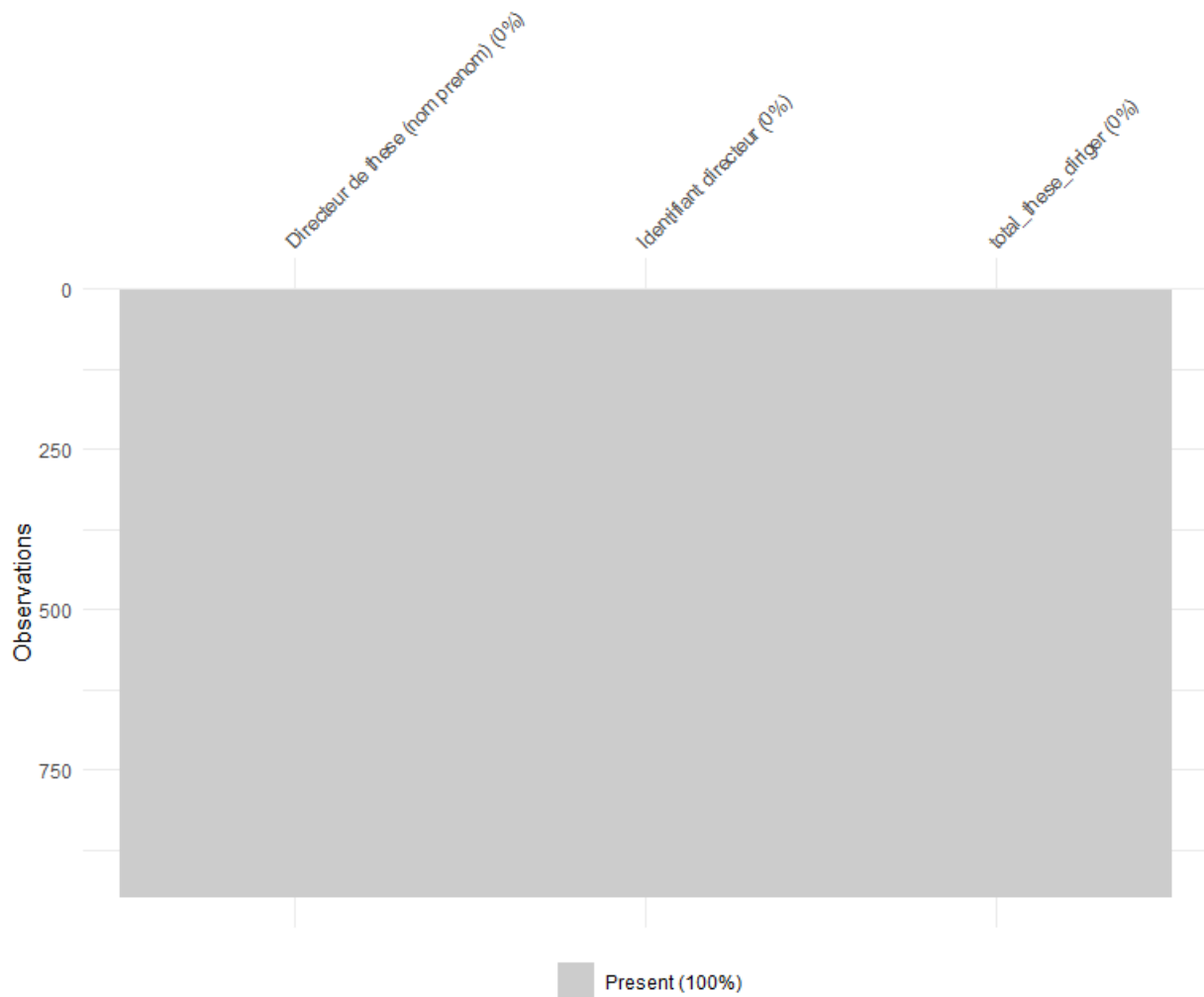
```
n_distinct(these_director_outliers_middle$`Identifiant directeur`)
```

```
## [1] 7
```

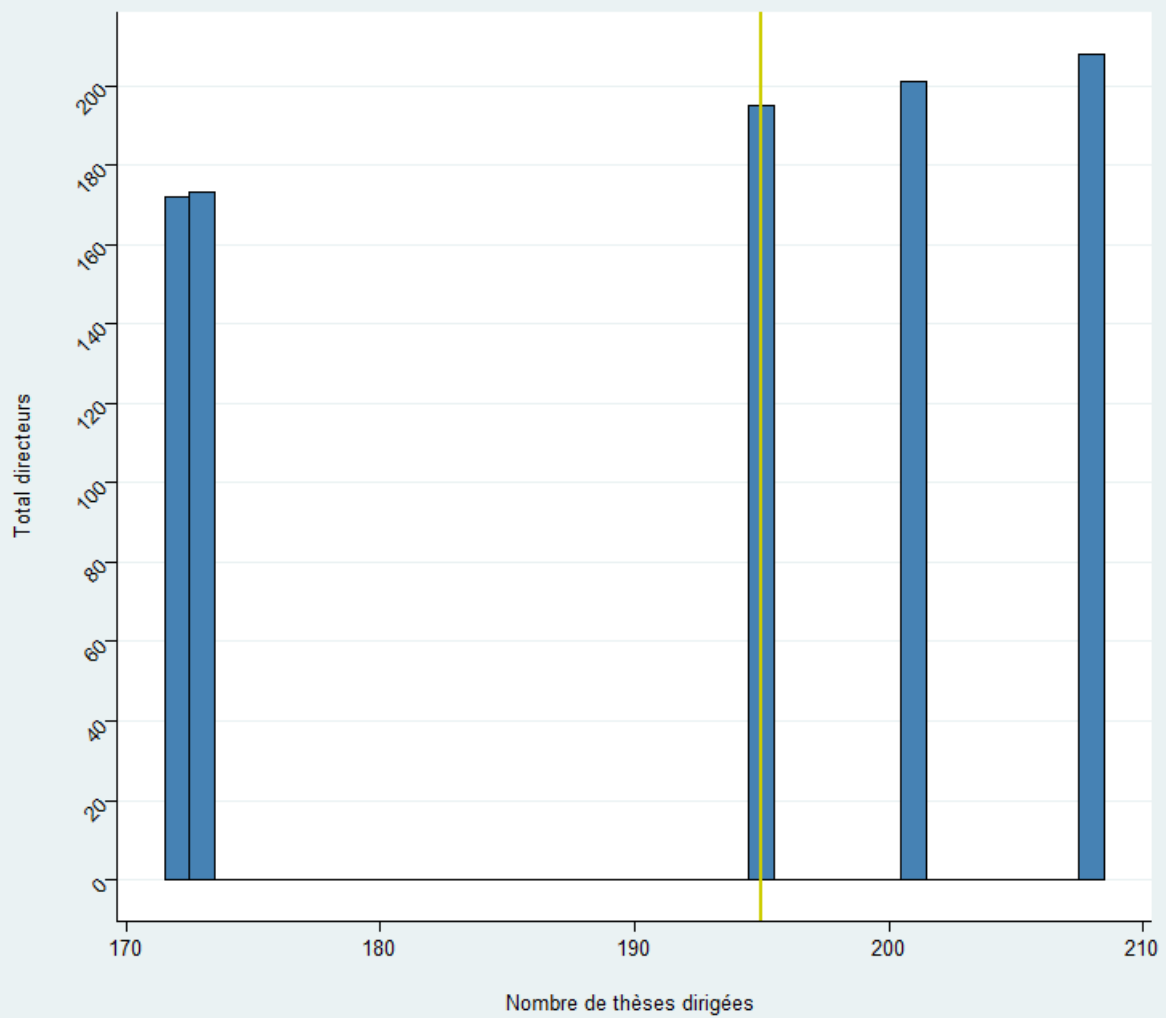
```
glimpse(these_director_outliers_middle)
```

```
## Rows: 949
## Columns: 3
## Groups: Directeur de these (nom prenom) [5]
## $ `Directeur de these (nom prenom)` <chr> "Blanc Francois-Paul", "Blanc Franco~
## $ `Identifiant directeur`      <chr> "26730774", "26730774", "26756625", ~
## $ total_these_diriger          <int> 201, 201, 195, 172, 195, 195, 1~
```

```
vis_miss(these_director_outliers_middle)
```



```
ggplot(these_director_outliers_middle, aes(total_these_diriger)) +  
  geom_histogram(binwidth = 1, fill = "steelblue", color = "Black") +  
  geom_vline(aes(xintercept = median(total_these_diriger)), color = "yellow3", size = 1) +  
  labs(x = "\nNombre de thèses dirigées ",  
       y = "Total directeurs\n") +  
  scale_x_continuous(breaks = seq(170, 210, 10)) +  
  scale_y_continuous(breaks = seq(0, 200, 20)) +  
  theme_stata() +  
  theme(axis.text.y = element_text(angle = 45))
```



```
# Total theses > 700.
these_director_outliers_big <- these_director %>% filter(total_these_diriger > 250)
dim(these_director_outliers_big)
```

```
## [1] 711 3
```

```
n_distinct(these_director_outliers_big$`Directeur de these (nom prenom)`)
```

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

```
unique(these_director_outliers_big$`Directeur de these (nom prenom)`)
```

```
## [1] "Directeur de these inconnu"
```

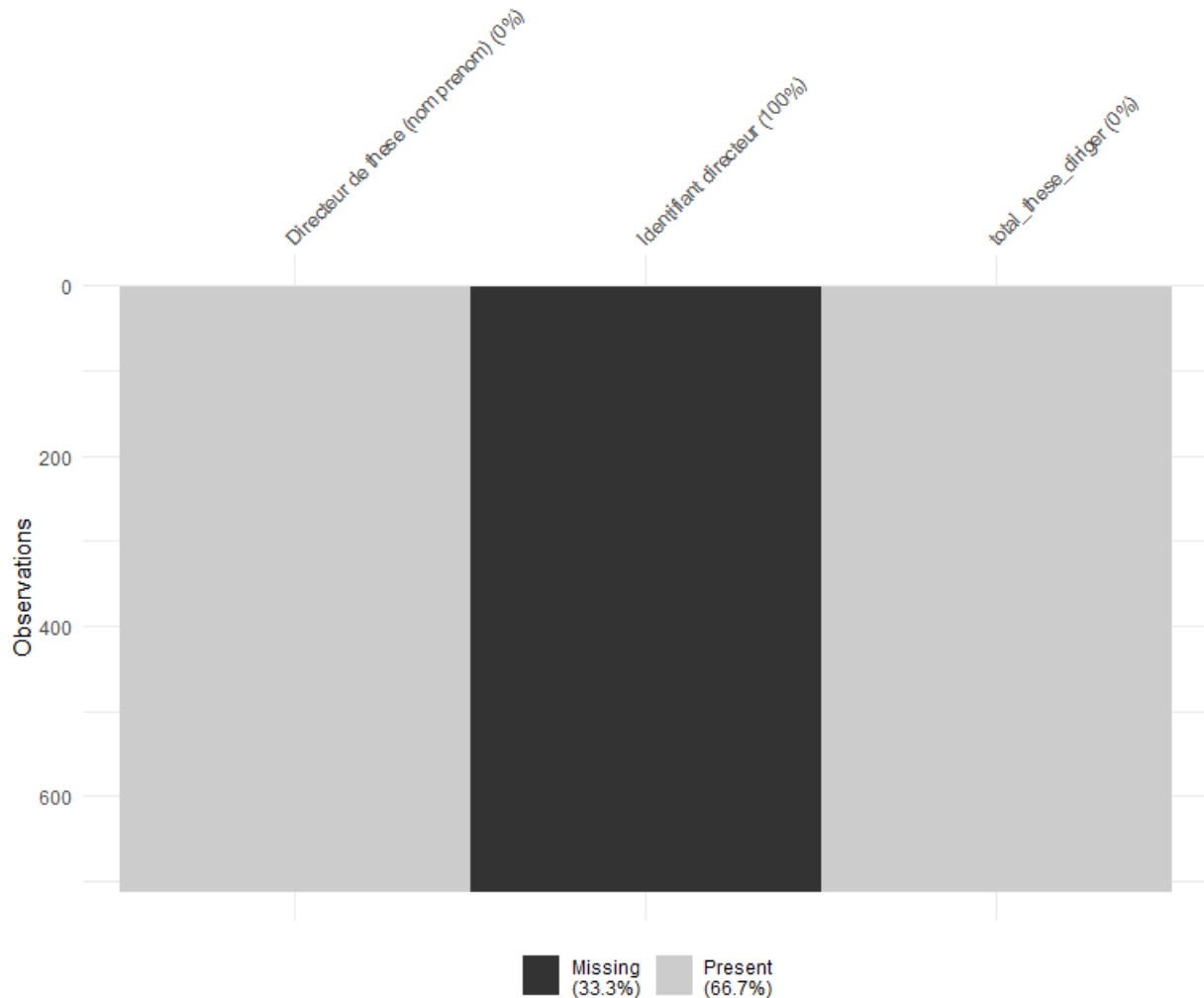
```
n_distinct(these_director_outliers_big$`Identifiant directeur`)
```

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

```
glimpse(these_director_outliers_big)
```

```
## Rows: 711
## Columns: 3
## Groups: Directeur de these (nom prenom) [1]
## $ `Directeur de these (nom prenom)` <chr> "Directeur de these inconnu", "Direc~
## $ `Identifiant directeur`      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ total_these_diriger          <int> 711, 711, 711, 711, 711, 711, 711, 7~
```

```
vis_miss(these_director_outliers_big)
```



5 Résultats préliminaires

5.1 Langues

5.1.1 Import et préparation des données

```
these_langue <- these
these_langue <- rename(these_langue, Langue = `Langue de la these`)
these_langue <- these_langue %>%
  mutate(Langue = as.factor(case_when(
    is.na(Langue) ~ "NA",
    Langue == "fr" ~ "Français",
    Langue == "en" ~ "Anglais",
    Langue == "enfr" | Langue == "fren" ~ "Bilingue",
    TRUE ~ "Autres")))
levels(these_langue$Langue)
```

```
## [1] "Anglais" "Autres" "Bilingue" "Français" "NA"
```


summary(these_langue)

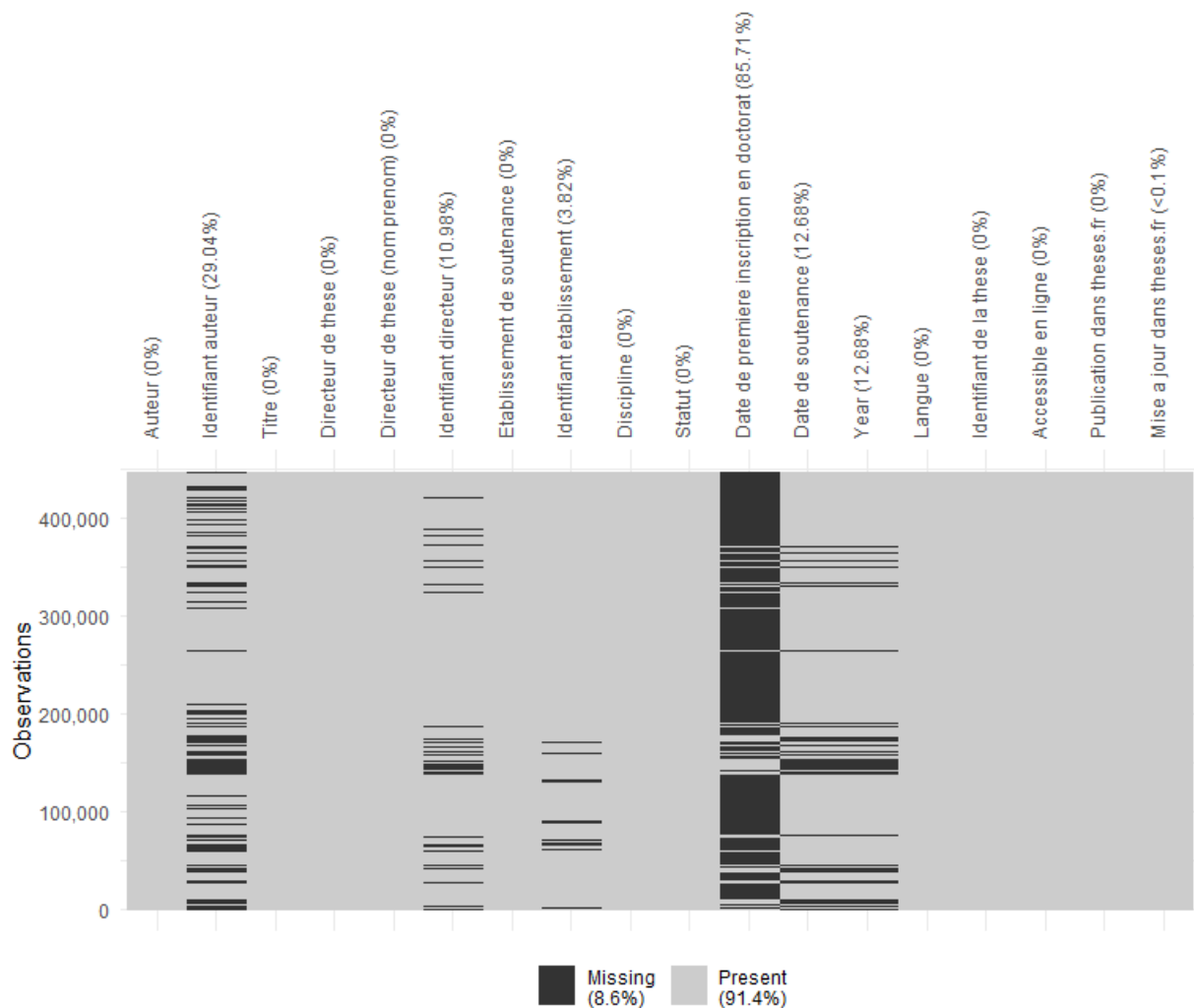
```
## Auteur      Identifiant auteur  Titre      Directeur de these
## Length:447644  Length:447644  Length:447644  Length:447644
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##
## Directeur de these (nom prenom) Identifiant directeur
## Length:447644      Length:447644
## Class :character    Class :character
## Mode  :character    Mode  :character
##
##
##
##
## Etablissement de soutenance Identifiant etablisement Discipline
## Length:447644      Length:447644      Length:447644
## Class :character    Class :character    Class :character
## Mode  :character    Mode  :character    Mode  :character
##
##
##
##
## Statut      Date de premiere inscription en doctorat
## enCours : 66329 Length:447644
## soutenue:381315 Class :character
##           Mode  :character
##
##
##
##
## Date de soutenance      Year      Langue      Identifiant de la these
## Min.  :1971-01-01  Min.  :1971  Anglais : 30942  Length:447644
## 1st Qu.:1994-01-01  1st Qu.:1994  Autres  : 3164  Class :character
## Median :2004-01-01  Median :2004  Bilingue: 15369  Mode  :character
## Mean   :2003-06-08  Mean   :2003  Français:334404
## 3rd Qu.:2012-06-25  3rd Qu.:2012  NA      : 63765
## Max.   :2020-07-07  Max.   :2020
## NA's   :56746      NA's   :56746
## Accessible en ligne Publication dans theses.fr Mise a jour dans theses.fr
## non:347341      Min.   :2006-04-13      Min.   :2010-10-12
## oui:100303      1st Qu.:2013-05-24      1st Qu.:2019-04-08
##           Median :2013-05-24      Median :2020-02-26
##           Mean   :2014-11-09      Mean   :2019-06-30
##           3rd Qu.:2016-07-11      3rd Qu.:2020-03-08
##           Max.   :2020-07-08      Max.   :2020-07-08
##           NA's   :177
```

```
these_langue %>%
  summarise(across(everything(), n_distinct)) %>%
  glimpse()
```

```
## Rows: 1
## Columns: 18
## $ Auteur <int> 430277
## $ `Identifiant auteur` <int> 313775
## $ Titre <int> 446815
## $ `Directeur de these` <int> 159019
## $ `Directeur de these (nom prenom)` <int> 159021
## $ `Identifiant directeur` <int> 98907
## $ `Etablissement de soutenance` <int> 568
## $ `Identifiant etablissement` <int> 573
## $ Discipline <int> 24263
## $ Statut <int> 2
## $ `Date de premiere inscription en doctorat` <int> 4010
## $ `Date de soutenance` <int> 3992
## $ Year <int> 45
## $ Langue <int> 5
## $ `Identifiant de la these` <int> 447572
## $ `Accessible en ligne` <int> 2
## $ `Publication dans theses.fr` <int> 2765
## $ `Mise a jour dans theses.fr` <int> 2634
```

5.1.2 Données manquantes

```
vis_miss(these_langue, warn_large_data = FALSE)+
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(angle = 90))
```



5.1.3 Exploration

```
# Fréquence des langues.
```

```
these_langue_count_year <- these_langue %>%
```

```
  select(Year) %>%
```

```
  count(Year) %>%
```

```
  rename(total_year = n)
```

```
these_langue_count_langue <- these_langue %>%
```

```
  count(Year, Langue) %>%
```

```
  rename(langue_count = n)
```

```
these_langue_fulljoin <- full_join(these_langue_count_year,
```

```
  these_langue_count_langue,
```

```
  by = "Year") %>%
```

```
  mutate(freq = langue_count / total_year)
```

```
# Jeu de données entier
```

```
these_langue_fulljoin %>%
```

```
ggplot(aes(Year, freq, color = Langue)) +
```

```
  geom_line(size = 1) +
```

```
  scale_x_continuous(breaks = seq(1970, 2020, 5)) +
```

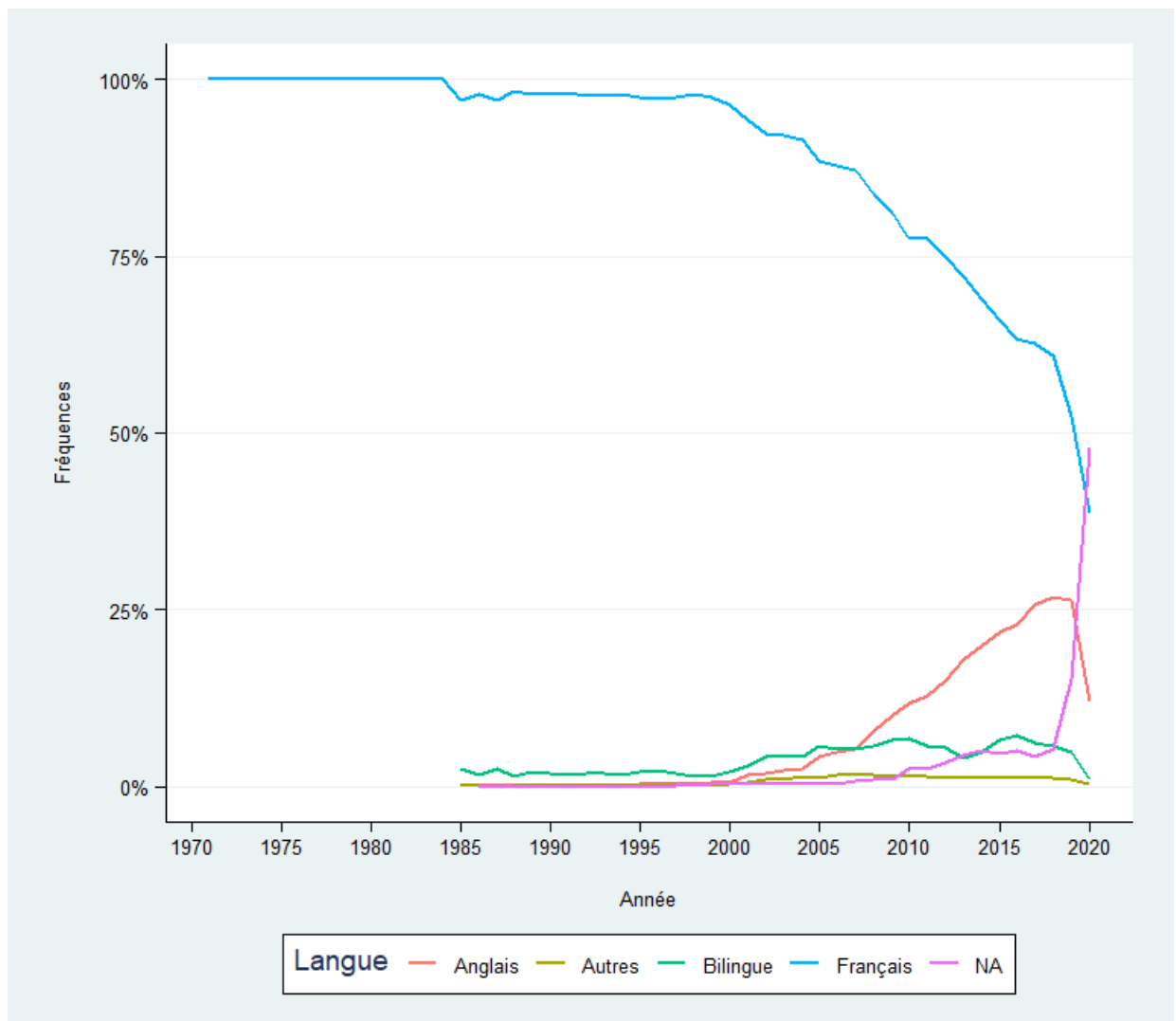
```
  scale_y_continuous(labels = percent_format()) +
```

```
  labs(x = "\nAnnée",
```

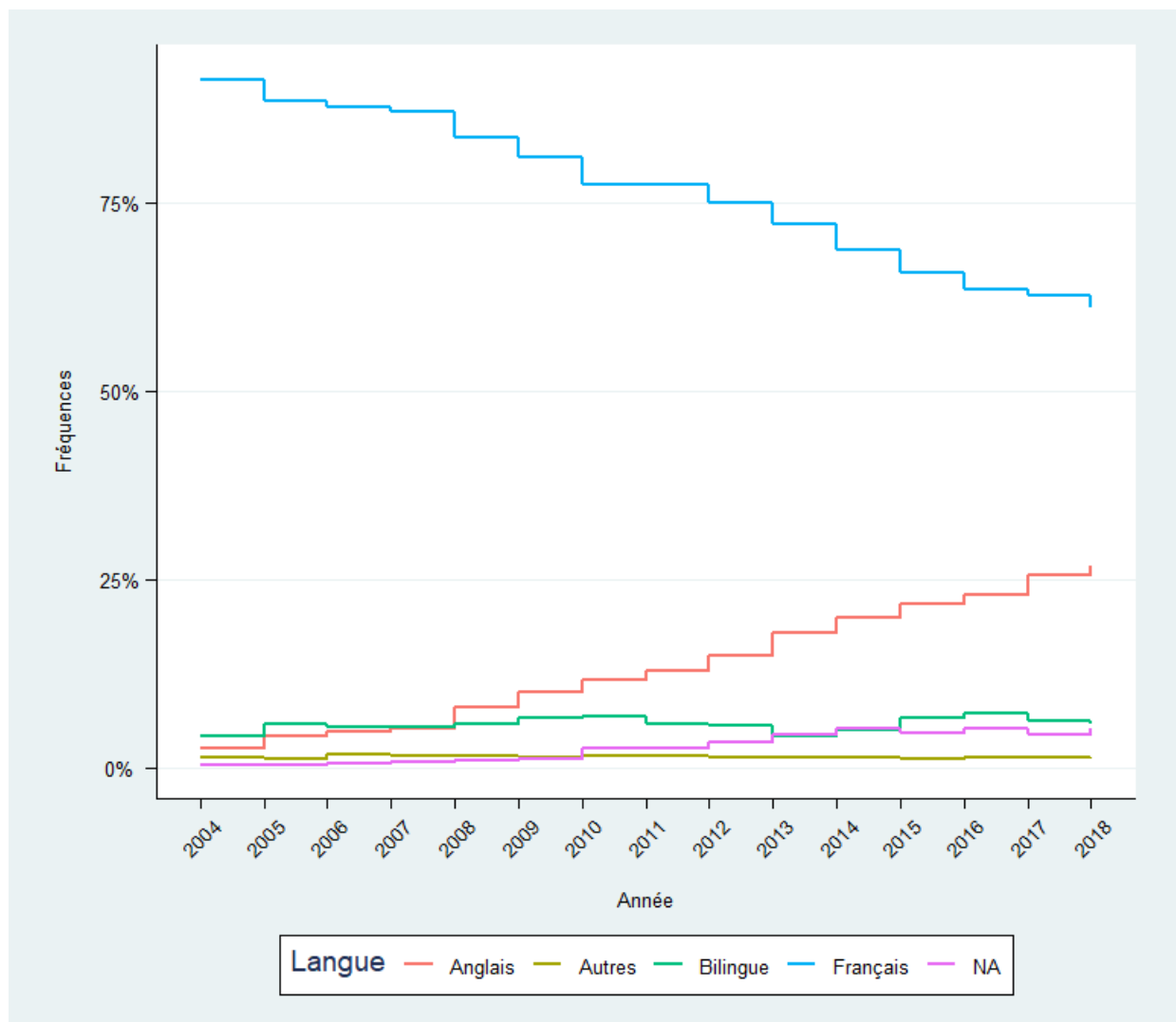
```
       y = "Fréquences\n") +
```

```
  theme_stata() +
```

```
  theme(axis.text.y = element_text(angle = 0))
```



```
# Période de 2004 à 2018
these_langue_fulljoin %>%
  filter(Year >= 2004 & Year <= 2018) %>%
  ggplot(aes(Year, freq, color = Langue)) +
  geom_step(size = 1) +
  scale_x_continuous(breaks = seq(2004, 2018, 1)) +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "\nAnnée",
       y = "Fréquences\n") +
  theme_stata() +
  theme(axis.text.x = element_text(angle = 45,
                                     vjust = 0.5),
        axis.text.y = element_text(angle = 0,
                                     hjust = 0.5))
```



6 SQL

```
business <- read_csv("jeux_de_donnees/business.csv")

glimpse(business)
```

```
## Rows: 4
## Columns: 4
## $ INCORP_DATE <date> 1995-05-01, 2001-01-01, 2002-06-30, 1999-05-01
## $ NAME      <chr> "Chilton Engineering", "Northeast Cooling Inc.", "Superio~
## $ STATE_ID  <chr> "12-345-678", "23-456-789", "34-567-890", "45-678-901"
## $ CUST_ID   <dbl> 10, 11, 12, 13
```

7 Travail en bonus

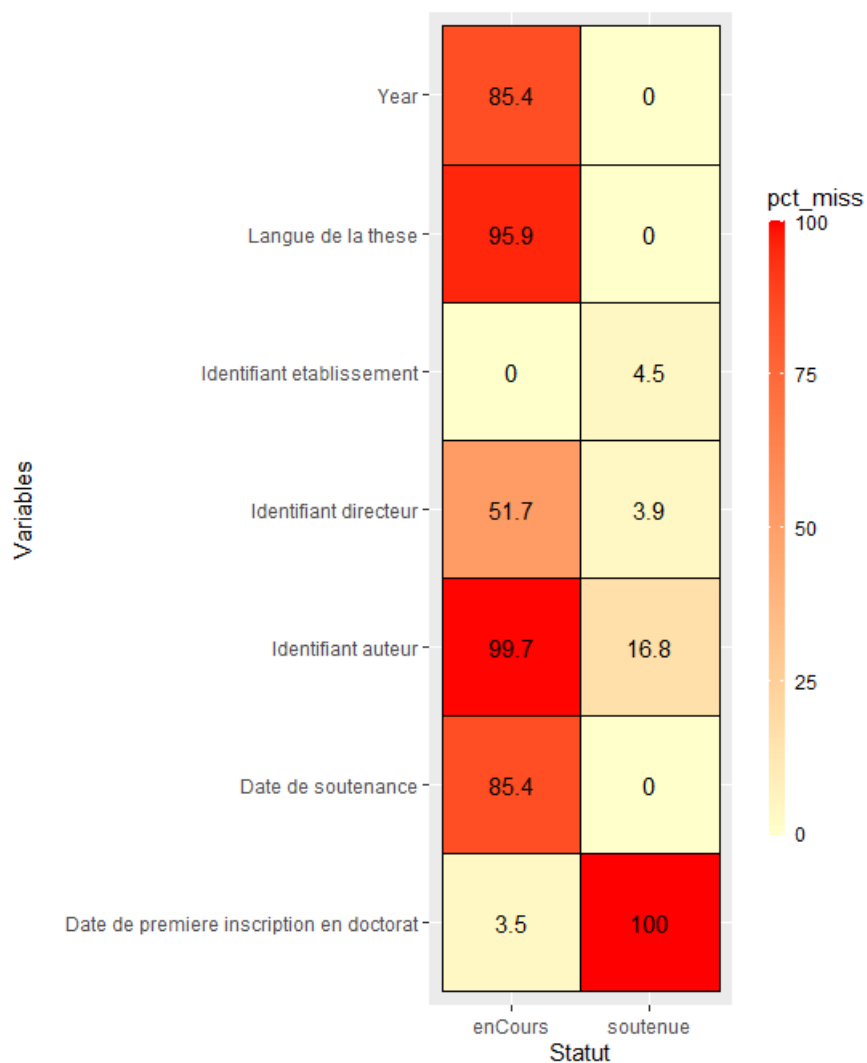
7.1 Heatmap des données manquantes

7.1.1 Import et préparation des données

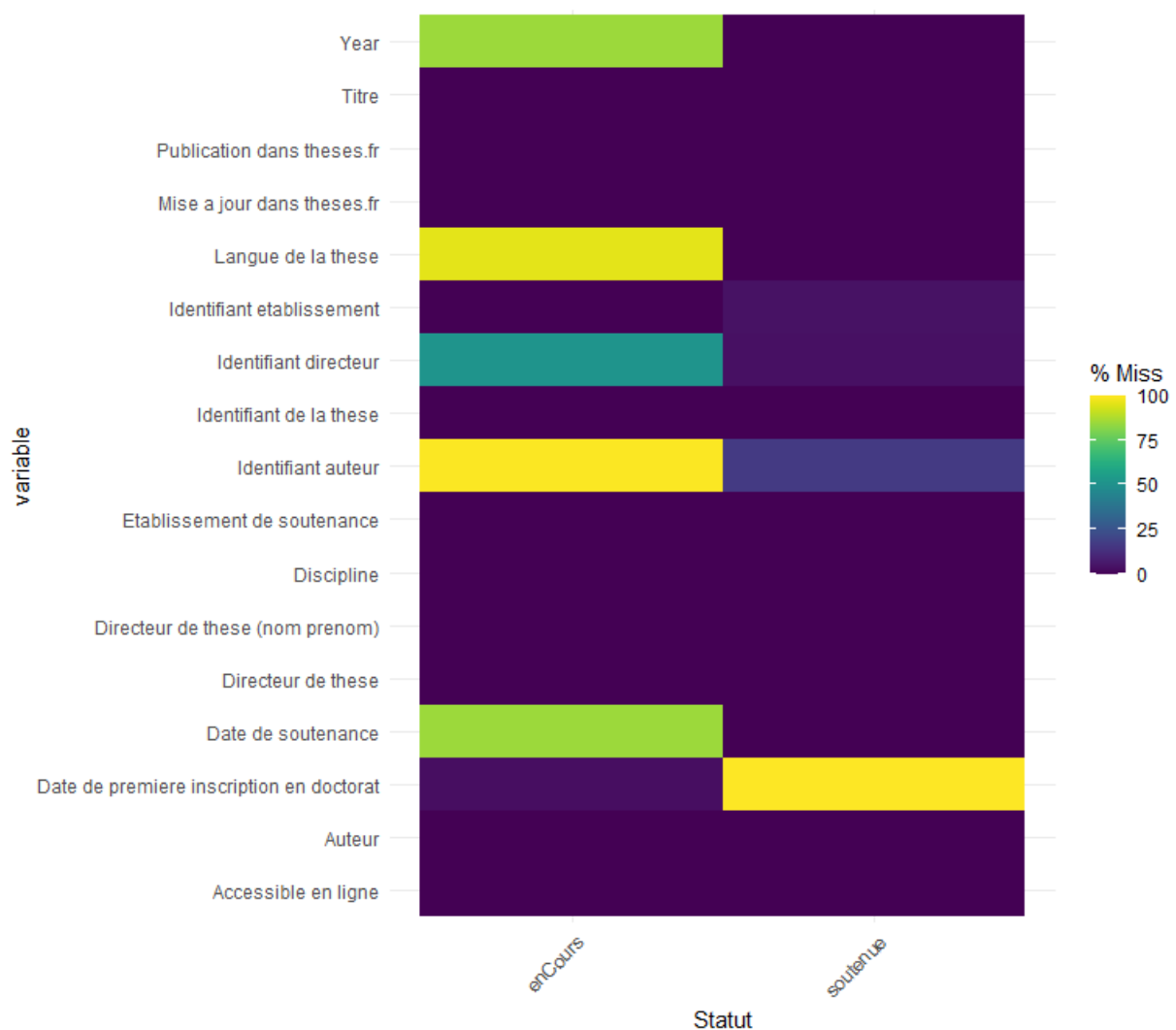
```
these_missing_heatmap <- these %>%
  select(Year, `Langue de la these`, `Identifiant etablissement`, `Identifiant directeur`,
    `Identifiant auteur`, `Date de soutenance`, `Date de premiere inscription en doctorat`,
    Statut) %>%
  group_by(Statut) %>%
  miss_var_summary()
these_missing_heatmap$pct_miss <- round(these_missing_heatmap$pct_miss, 1)
```

7.1.2 Visualisations

```
these_missing_heatmap %>%
  ggplot(aes(Statut, variable, fill = pct_miss)) +
  geom_tile(color = "black") +
  geom_text(aes(label = pct_miss), color = "black", size = 4) +
  labs(y = "Variables\n") +
  scale_fill_gradient2(low = "#075AFF",
    mid = "#FFFFCC",
    high = "#FF0000") +
  guides(fill = guide_colourbar(barwidth = 0.5,
    barheight = 20)) +
  coord_fixed()
```



```
gg_miss_fct(x = these, fct = Statut)
```



Les deux heatmap représente la même chose mais montre deux moyens de le faire différemment.

7.2 Problème Genre/Discipline/Langue

7.2.1 Problème Genre/Discipline import et préparation des données PhD_v2_gender

```
these_gender <- as_tibble(read_csv("jeux_de_donnees/PhD_v2_gender.csv"))
these_gender <- subset(these_gender, select = -c(...1))
these_gender$`Date de premiere inscription en doctorat` <- dmy(these_gender$`Date de premiere inscription en doctorat`)
these_gender$`Date de soutenance` <- dmy(these_gender$`Date de soutenance`)
these_gender$`Publication dans theses.fr` <- dmy(these_gender$`Publication dans theses.fr`)
these_gender$`Mise a jour dans theses.fr` <- dmy(these_gender$`Mise a jour dans theses.fr`)
these_gender$Statut <- as.factor(these_gender$Statut)
these_gender$`Langue de la these` <- as.factor(these_gender$`Langue de la these`)
these_gender$`Accessible en ligne` <- as.factor(these_gender$`Accessible en ligne`)
these_gender$`Identifiant directeur` <- na_if(these_gender$`Identifiant directeur`, "na")
these_gender$gender <- as.factor(these_gender$gender)
these_gender$Gender <- these_gender$gender

these_gender %>%
  subset(select = -c(gender)) %>%
  summarise(across(everything(), n_distinct)) %>%
  glimpse()
```

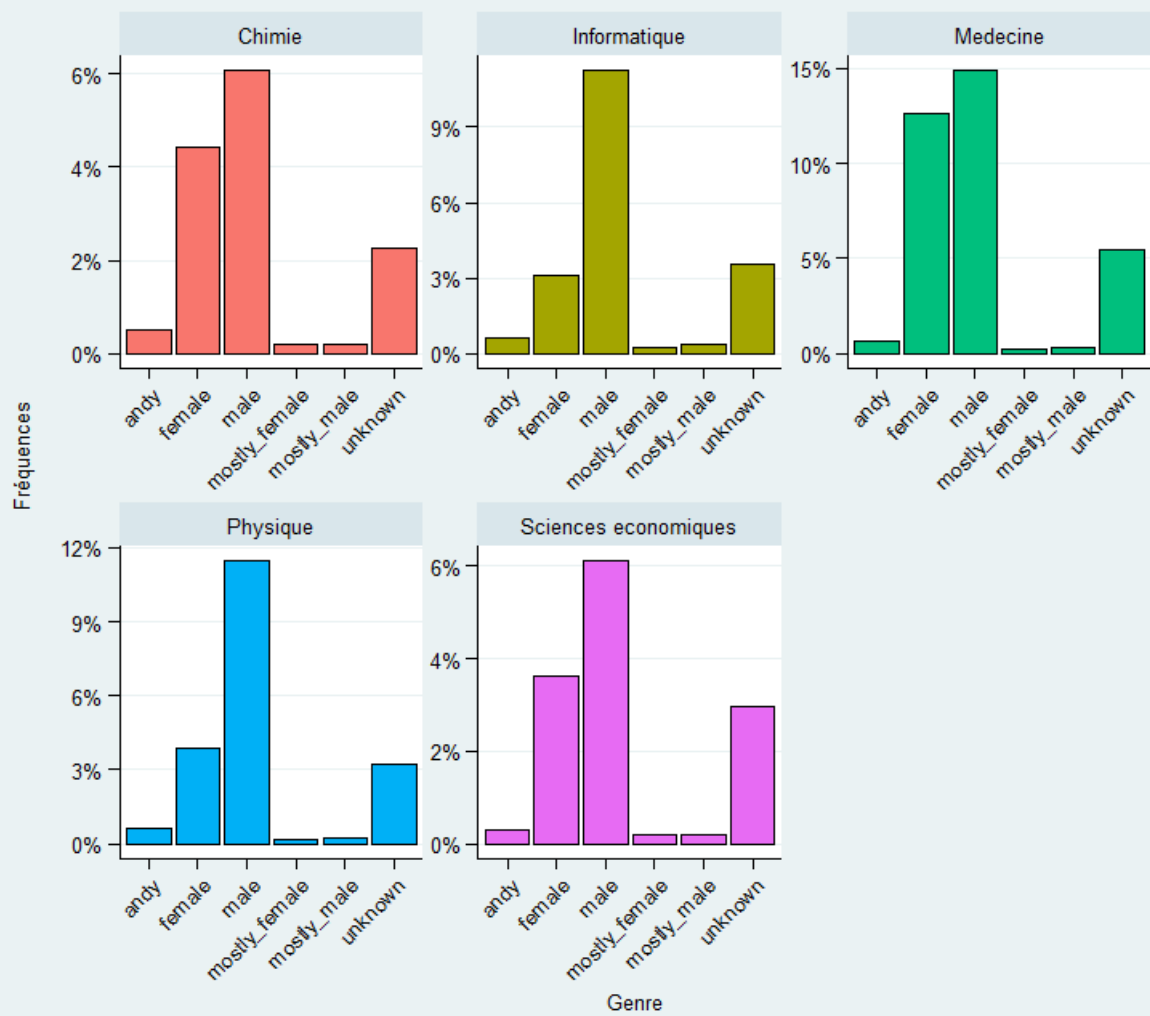
```
## Rows: 1
## Columns: 21
## $ Auteur <int> 430277
## $ `Identifiant auteur` <int> 313775
## $ Titre <int> 446815
## $ `Directeur de these` <int> 159019
## $ `Directeur de these (nom prenom)` <int> 159021
## $ `Identifiant directeur` <int> 98907
## $ `Etablissement de soutenance` <int> 568
## $ `Identifiant etablissement` <int> 573
## $ Discipline <int> 24263
## $ Statut <int> 2
## $ `Date de premiere inscription en doctorat` <int> 4010
## $ `Date de soutenance` <int> 3992
## $ Year <int> 45
## $ `Langue de la these` <int> 206
## $ `Identifiant de la these` <int> 447572
## $ `Accessible en ligne` <int> 2
## $ `Publication dans theses.fr` <int> 2765
## $ `Mise a jour dans theses.fr` <int> 2634
## $ First_name <int> 44920
## $ Last_name <int> 237706
## $ Gender <int> 6
```

7.2.2 Visualisations

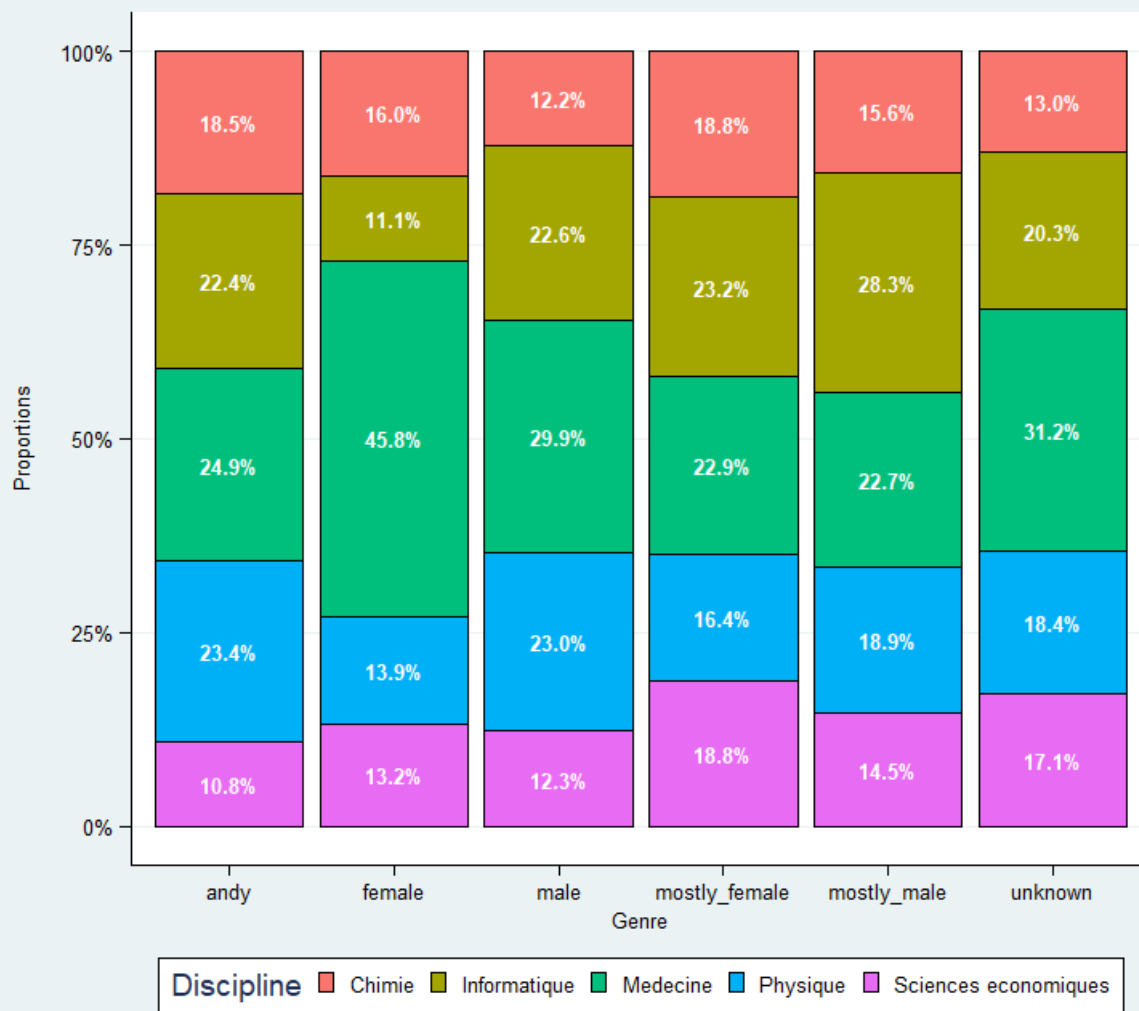
```
# Selection des top 5 disciplines
these_gender_top_5_discipline <- these_gender %>%
  select(Discipline, Gender) %>%
  count(Discipline, sort = TRUE) %>%
  slice(1:5) %>%
  subset(select = -c(n)) %>%
  pull()

these_gender_top_5_discipline <- these_gender %>%
  filter(Discipline %in% these_gender_top_5_discipline)

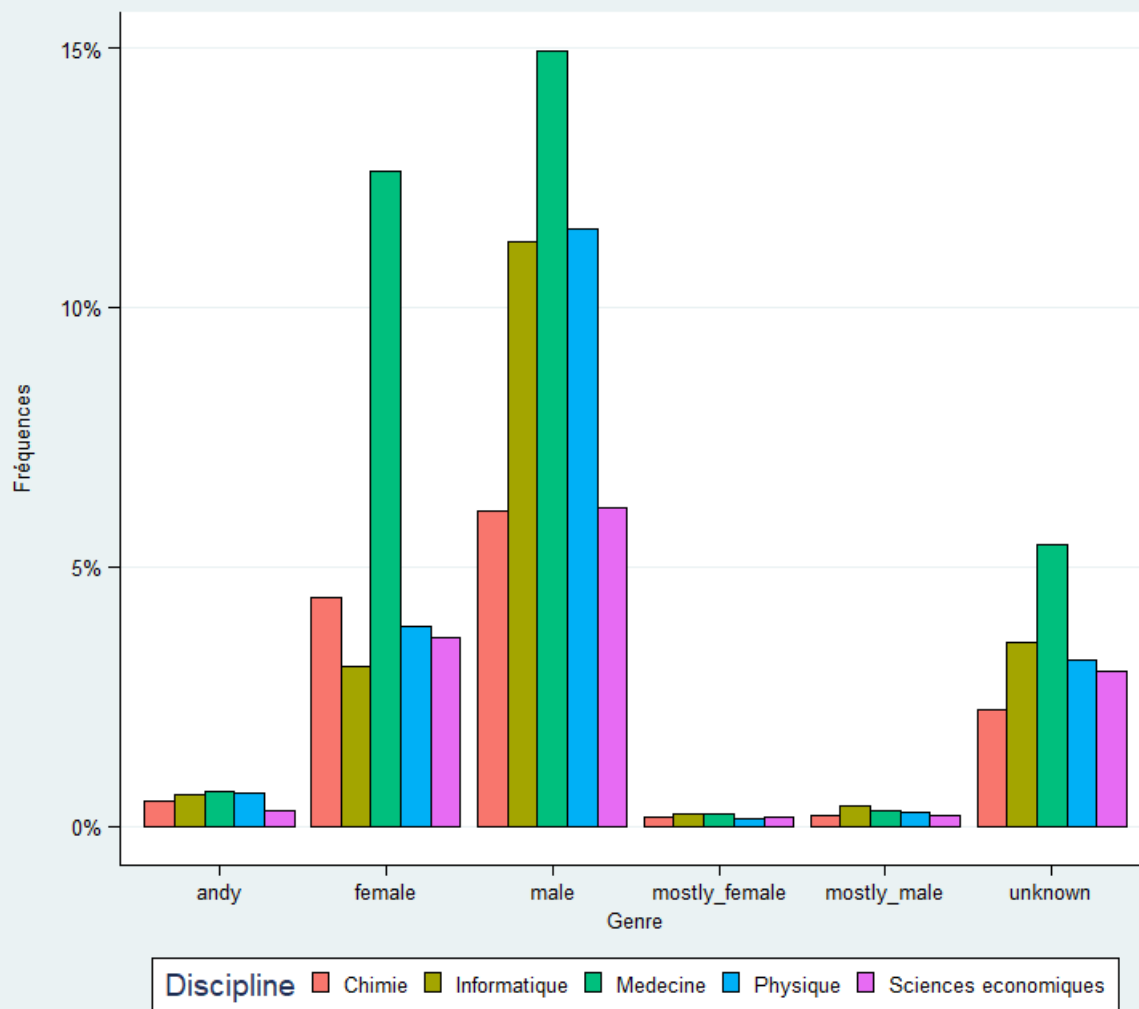
# Visualisations
these_gender_top_5_discipline %>%
  ggplot(aes(Gender, after_stat(count/sum(count)), fill = Discipline)) +
  geom_bar( position = "dodge", color = "black") +
  facet_wrap( ~ Discipline, scales = "free") +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "Genre",
       y = "Fréquences\n") +
  theme_stata() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1),
        strip.text.x = element_text(size = 10),
        axis.text.y = element_text(angle = 0))
```



```
these_gender_top_5_discipline %>%
  ggplot(aes(Gender, after_stat(count/sum(count)), fill = Discipline, by = Gender)) +
  geom_bar(position = "fill", color = "black") +
  geom_text(stat = "prop", position = position_fill(.5),
    colour = "white", fontface = "bold", size = 3.5) +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "Genre",
    y = "Proportions\n") +
  theme_stata() +
  theme(legend.position = "bottom",
    legend.key.size = unit(0.3, 'cm'),
    legend.key.height = unit(0.3, 'cm'),
    legend.key.width = unit(0.3, 'cm'),
    legend.title = element_text(size=14),
    legend.text = element_text(size=10),
    axis.text.y = element_text(angle = 0))
```

```
these_gender_top_5_discipline %>%
  ggplot(aes(Gender, after_stat(count/sum(count)), fill = Discipline)) +
  geom_bar(position = "dodge", color = "black") +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "Genre",
       y = "Fréquences\n") +
  theme_stata() +
  theme(legend.position = "bottom",
        legend.key.size = unit(0.3, 'cm'),
        legend.key.height = unit(0.3, 'cm'),
        legend.key.width = unit(0.3, 'cm'),
        legend.title = element_text(size=14),
        legend.text = element_text(size=10),
        axis.text.y = element_text(angle = 0))
```



7.2.3 Problème Genre/Discipline import et préparation des données PhD_v3

```
these_v3 <- as_tibble(read_csv("jeux_de_donnees/PhD_v3.csv"))
these_v3 <- subset(these_v3, select = -c(...1))
these_v3$`Date de premiere inscription en doctorat` <- dmy(these_v3$`Date de premiere inscription en doctorat`)
these_v3$`Date de soutenance` <- dmy(these_v3$`Date de soutenance`)
these_v3$`Publication dans theses.fr` <- dmy(these_v3$`Publication dans theses.fr`)
these_v3$`Mise a jour dans theses.fr` <- dmy(these_v3$`Mise a jour dans theses.fr`)
these_v3$Statut <- as.factor(these_v3$Statut)
these_v3$`Langue de la these` <- as.factor(these_v3$`Langue de la these`)
these_v3$`Accessible en ligne` <- as.factor(these_v3$`Accessible en ligne`)
these_v3$`Identifiant directeur` <- na_if(these_v3$`Identifiant directeur`, "na")
these_v3$Genre <- as.factor(these_v3$Genre)
these_v3 <- rename(these_v3, Discipline_prediction = `Discipline_prÃ©di`)
these_v3$Discipline_prediction <- as.factor(these_v3$Discipline_prediction)
levels(these_v3$Discipline_prediction)[levels(these_v3$Discipline_prediction) == "MathÃ©matiques"] <- "Mathematiques"
levels(these_v3$Discipline_prediction)[levels(these_v3$Discipline_prediction) == "Science de l'ingÃ©nieur"] <- "Science de l'ingenieur"

these_v3 %>%
  summarise(across(everything(), n_distinct)) %>%
  glimpse()
```

```
## Rows: 1
## Columns: 22
## $ Auteur <int> 430272
## $ `Identifiant auteur` <int> 313772
## $ Titre <int> 446812
## $ `Directeur de these` <int> 159018
## $ `Directeur de these (nom prenom)` <int> 159020
## $ `Identifiant directeur` <int> 98906
## $ `Etablissement de soutenance` <int> 568
## $ `Identifiant etablissement` <int> 573
## $ Discipline <int> 24262
## $ Statut <int> 2
## $ `Date de premiere inscription en doctorat` <int> 4010
## $ `Date de soutenance` <int> 3992
## $ Year <int> 45
## $ `Langue de la these` <int> 206
## $ `Identifiant de la these` <int> 447567
## $ `Accessible en ligne` <int> 2
## $ `Publication dans theses.fr` <int> 2765
## $ `Mise a jour dans theses.fr` <int> 2634
## $ Discipline_prediction <int> 15
## $ Genre <int> 6
## $ etablissement_rec <int> 111
## $ Langue_rec <int> 5
```

7.2.4 Préparation des tables intermédiaires

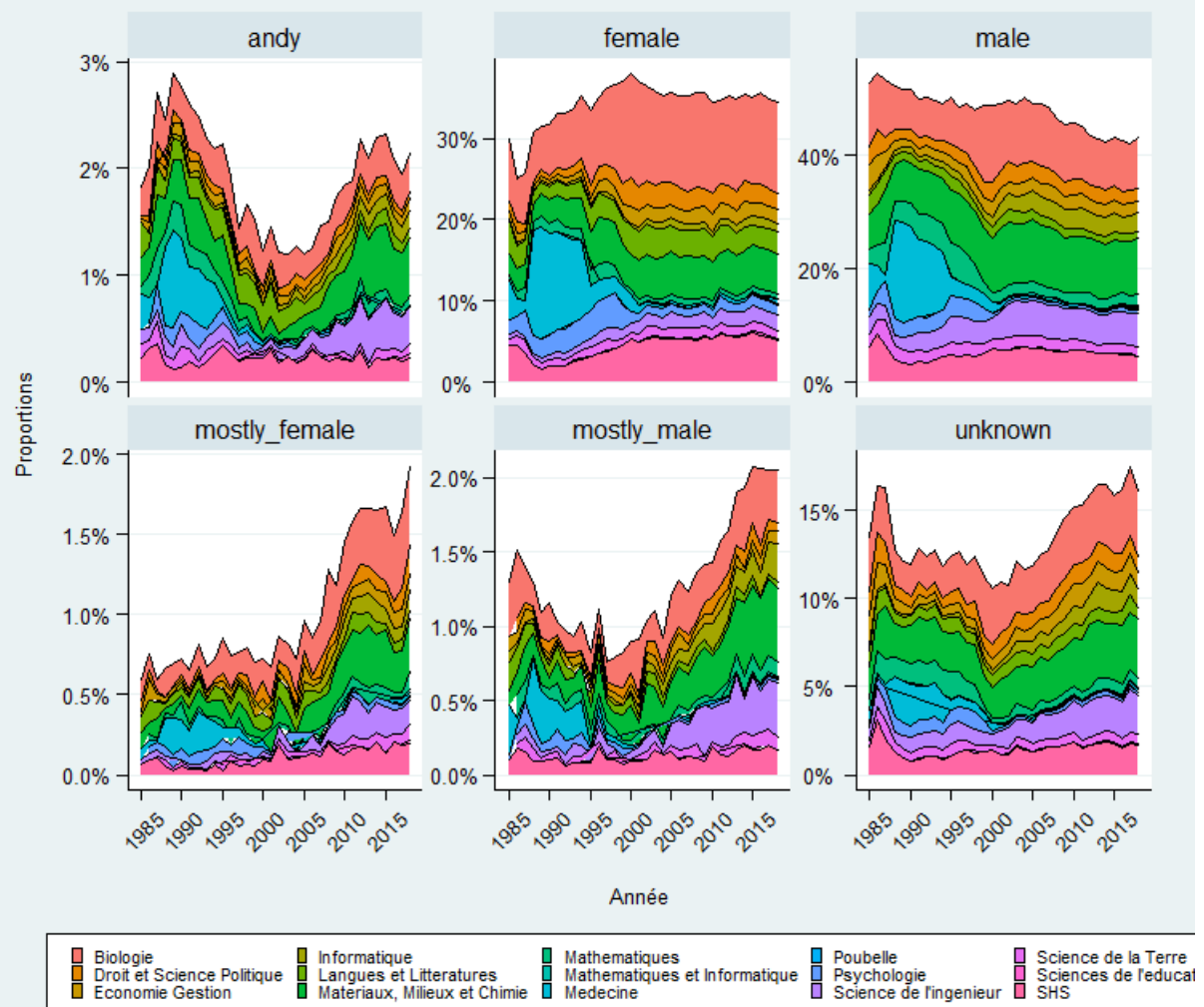
```
these_v3_count_year <- these_v3 %>%
  filter(Year >= 1985 & Year <= 2018) %>%
  count(Year) %>%
  rename(total_year = n)

these_v3_year_genre_discipline <- these_v3 %>%
  filter(Year >= 1985 & Year <= 2018) %>%
  count(Year, Genre, Discipline_prediction) %>%
  rename(total = n)

these_genre_discipline_full <- full_join(these_v3_count_year,
  these_v3_year_genre_discipline,
  by = "Year") %>%
  mutate(freq = total / total_year) %>%
  drop_na()
```

7.2.5 Visualisation

```
these_genre_discipline_full %>%
  ggplot(aes(Year, freq, fill = Discipline_prediction)) +
  geom_area(color = "black") +
  facet_wrap( ~ Genre, scales = "free_y") +
  scale_x_continuous(breaks = seq(1985, 2020, 5)) +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "\nAnnée",
    y = "Proportions\n") +
  theme_stata() +
  theme(axis.text.x = element_text(angle = 45,
    vjust = 0.5),
    axis.text.y = element_text(angle = 0,
    hjust = 0.5),
    legend.key.size = unit(0.2, "cm"),
    legend.text = element_text(size = '8'),
    legend.title = element_blank())
```



7.2.6 Problème Langue/Discipline import et préparation des données PhD_v3

```
these_v3_langue <- these_v3
these_v3_langue <- rename(these_v3_langue, Langue = `Langue de la these`)
these_v3_langue <- these_v3_langue %>%
  mutate(Langue = as.factor(case_when(
    is.na(Langue) ~ "NA",
    Langue == "fr" ~ "Français",
    Langue == "en" ~ "Anglais",
    Langue == "enfr" | Langue == "fren" ~ "Bilingue",
    TRUE ~ "Autres")))
```

7.2.7 Préparation des tables intermédiaires

```
these_soutenance_year_langue_discipline <- these_v3_langue %>%
  filter(Year >= 1985 & Year <= 2018) %>%
  count(Year, Langue, Discipline_prediction) %>%
  rename(total = n)

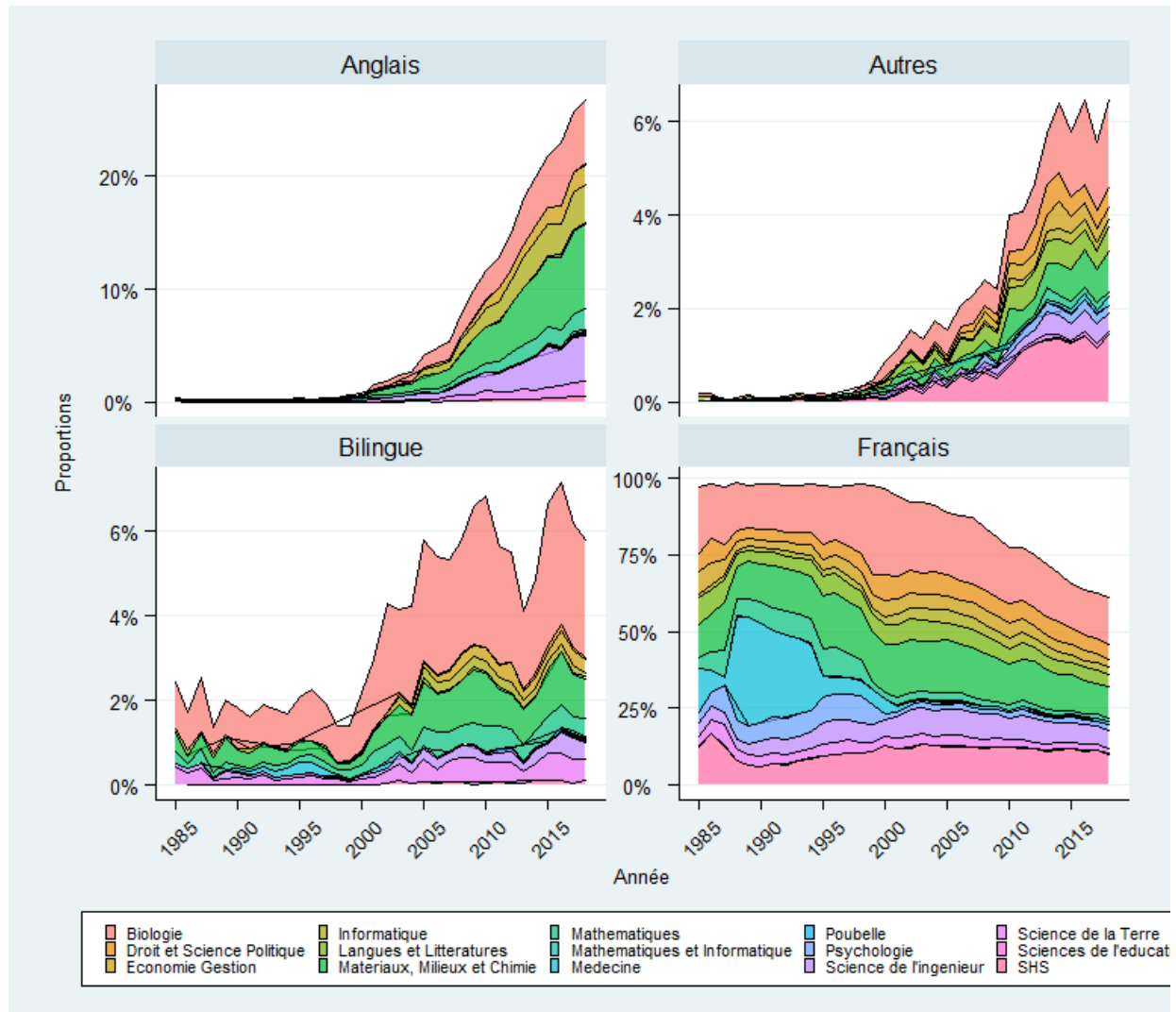
these_genre_discipline_full <- full_join(these_v3_count_year,
  these_soutenance_year_langue_discipline,
  by = "Year") %>%
  mutate(freq = total / total_year)
```

7.2.8 Visualisation

```

these_genre_discipline_full %>%
  ggplot(aes(Year, freq, fill = Discipline_prediction)) +
  geom_area(color = "black", alpha = 0.7) +
  facet_wrap(~ Langue, scales = "free_y") +
  scale_x_continuous(breaks = seq(1985, 2020, 5)) +
  scale_y_continuous(labels = percent_format()) +
  labs(x = "Année",
       y = "Proportions\n") +
  theme_stata() +
  theme(axis.text.x = element_text(angle = 45,
                                     vjust = 0.5),
        axis.text.y = element_text(angle = 0,
                                     hjust = 0.5),
        legend.key.size = unit(0.2, "cm"),
        legend.text = element_text(size = '8'),
        legend.title = element_blank())

```



7.3 Webscraping

Déclaration des séquences

```

langue <- c("fr", "en", "es")
numero_page <- seq(0, 500, 10)
numero_page_langue <- seq(0, 300, 10)

```

Déclaration des tibbles

```

thesis_webscraping_info <- tibble()
thesis_webscraping_pdf <- tibble()
thesis_webscraping_langue <- tibble()

```

Scrap des informations

```

for (page_result in numero_page) {
  link_infos <- paste0("https://theses.fr/fr/?q=&fq=dateSoutenance:[1965-01-01T23:59:59Z%2BTO%2B2022-12-31T23:59:59Z]&checkedfacets=&start=",
    page_result, "&sort=none&status=&access=&prevision=&filtrepersonne=&zone1=titreBAS&val1=&op1=AND&zone2=auteurs&val2=&op2=AND")
}

```

```

pages_infos <- read_html(link_infos)

Titre <- pages_infos %>% html_nodes("h2 a") %>% html_text2() %>% str_to_lower() %>%
  str_to_sentence()

Auteur <- pages_infos %>% html_nodes("#resultat p") %>% html_text2() %>%
  str_split("\\r", n = 2) %>% map_chr(1) %>% sub(".*? ", "", .)

Discipline <- pages_infos %>% html_nodes(".domaine h5:nth-child(1)") %>% html_text2()

Directeur <- pages_infos %>% html_nodes("#resultat p") %>% html_text2() %>%
  str_split("\\r", n = 3) %>%
  map_chr(3) %>% sub(".*? ", "", .) %>% str_replace(" et de \\r", ",") %>%
  str_split("\\r") %>% map_chr(1)

Etablissement <- pages_infos %>% html_nodes("#resultat p") %>% html_text2() %>%
  str_split("\\r", n = 3) %>%
  map_chr(3) %>% sub(".*? ", "", .) %>% str_replace(" et de \\r", ",") %>%
  str_split("\\r") %>% map_chr(2) %>% str_remove(" - ")

Statut <- pages_infos %>% html_nodes("div.statusThese img") %>% html_attr("title")

Date_soutenance_ymd <- pages_infos %>% html_nodes("br+ h5") %>% html_text2() %>%
  str_remove("En préparation depuis le ") %>% str_sub(13, 22) %>% dmy()

Date_soutenance_y <- pages_infos %>% html_nodes("br+ h5") %>% html_text2() %>%
  str_remove_all("En préparation depuis le ") %>% str_remove_all("Soutenue le ") %>%
  str_sub(12, 16) %>% as.Date(as.character(), format = "%Y") %>% year()

Date_inscription <- pages_infos %>% html_nodes("br+ h5") %>% html_text2() %>%
  str_sub(25, 35) %>% dmy()

thesis_webscraping_info <- rbind(thesis_webscraping_info, tibble(Auteur, Titre, Discipline, Directeur,
  Etablissement, Statut, Date_inscription,
  Date_soutenance_ymd, Date_soutenance_y))

}

##### Scrap des pdf #####

for (page_pdf in numero_page) {
  link_pdf <- paste0("https://theses.fr/fr/?q=&fq=dateSoutenance:[1965-01-01T23:59:59Z%2BTO%2B2022-12-31T23:59:59Z]&checkedfacets=&start=",page_pdf)
  pages_pdf <- read_html(link_pdf)

  Link_to_pdf <- pages_pdf %>% html_nodes(".arrondi-10x a") %>% html_attr("href") %>%
    paste0("theses.fr", .)

  Auteur <- pages_pdf %>% html_nodes("#resultat p") %>% html_text2() %>%
    str_split("\\r", n = 2) %>% map_chr(1) %>% sub(".*? ", "", .)

  thesis_webscraping_pdf <- rbind(thesis_webscraping_pdf, tibble(Auteur, Link_to_pdf))

}

##### Scrap des langues (Pbm duplicate later) #####

for(Lg in langue){
  for (npl in numero_page_langue) {
    link_langue <- paste0("https://theses.fr/fr/?q=&fq=dateSoutenance:[1965-01-01T23:59:59Z%2BTO%2B2022-12-31T23:59:59Z]&checkedfacets=langue")
    pages_langue <- read_html(link_langue)

    Auteur <- pages_langue %>% html_nodes("#resultat p") %>% html_text2() %>%
      str_split("\\r", n = 2) %>% map_chr(1) %>% sub(".*? ", "", .)

    Langue <- Lg

    thesis_webscraping_langue <- rbind(thesis_webscraping_langue, tibble(Auteur, Langue))

  }
}

```

```
##### Fusion des tibbles #####
```

```
thesis_webscraping <- left_join(thesis_webscraping_info, thesis_webscraping_pdf, by = "Auteur")
```

```
thesis_webscraping <- left_join(thesis_webscraping, thesis_webscraping_langue, by = "Auteur")
```

```
##### Data wrangling #####
```

```
thesis_webscraping <- thesis_webscraping %>% mutate(Statut = case_when(Statut == " thèse en cours de préparation" ~ "en_cours", Statut == "thèse s
```

```
glimpse(thesis_webscraping)
```

```
## Rows: 519
## Columns: 11
## $ Auteur      <chr> "Matthias Pautard ", "Mohamed Amine Khelif ", "Dam~
## $ Titre      <chr> "Simulation de la balistique intérieure de mortier~
## $ Discipline  <chr> "Énergétique", "STIC (Sciences et Technologies de ~
## $ Directeur   <chr> "Christian Chauveau", "Olivier Romain", "Colette D~
## $ Etablissement <chr> "Orléans ", "CY Cergy Paris Université ", "Valenci~
## $ Statut      <chr> "en_cours", "en_cours", "en_cours", "en_cours", "e~
## $ Date_inscription <date> 2020-11-09, NA, 2022-04-28, 2022-04-28, 2018-12-0~
## $ Date_soutenance_ymd <date> NA, 2021-07-12, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ Date_soutenance_y <dbl> NA, NA, NA, NA, NA, 2009, 2006, 2003, 2006, 2009, ~
## $ Link_to_pdf  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "theses.fr~
## $ Langue      <chr> NA, "fr", NA, NA, NA, "fr", "fr", "fr", NA, ~
```

```
summary(thesis_webscraping)
```

```
##   Auteur      Titre      Discipline      Directeur
## Length:519   Length:519   Length:519   Length:519
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
##
##
##
## Etablissement      Statut      Date_inscription
## Length:519         Length:519   Min. :2007-10-01
## Class :character   Class :character 1st Qu.:2021-07-13
## Mode :character    Mode :character Median :2021-10-01
##                    Mean :2021-03-25
##                    3rd Qu.:2021-12-28
##                    Max. :2022-05-16
##                    NA's :193
## Date_soutenance_ymd Date_soutenance_y Link_to_pdf      Langue
## Min. :2013-07-08 Min. :1985   Length:519      Length:519
## 1st Qu.:2021-06-21 1st Qu.:1986   Class :character Class :character
## Median :2021-12-14 Median :1997   Mode :character Mode :character
## Mean :2021-05-14 Mean :1998
## 3rd Qu.:2022-02-18 3rd Qu.:2008
## Max. :2022-04-19 Max. :2012
## NA's :435      NA's :410
```

```
head(thesis_webscraping)
```

```
## # A tibble: 6 x 11
##   Auteur      Titre Discipline Directeur Etablissement Statut Date_inscription
##   <chr>      <chr> <chr>    <chr>    <chr>    <chr>    <date>
## 1 "Matthias Pau~ Simu~ Energétiq~ "Christi~ "Orléans " en_co~ 2020-11-09
## 2 "Mohamed Amin~ Arch~ STIC (Sci~ "Olivier~ "CY Cergy P~ en_co~ NA
## 3 "Damien Galan~ Etud~ Mathemati~ "Colette~ "Valencienn~ en_co~ 2022-04-28
## 4 "Zikriya Khan~ Synt~ Electroni~ "Denis R~ "Valencienn~ en_co~ 2022-04-28
## 5 "Meiting Tu " Opti~ Automatiq~ "Dominiq~ " Olivier O~ en_co~ 2018-12-01
## 6 "Cécile Halté" Déte~ Chimie - ~ "Thierry~ "Grenoble 1~ soute~ NA
## # ... with 4 more variables: Date_soutenance_ymd <date>,
## #   Date_soutenance_y <dbl>, Link_to_pdf <chr>, Langue <chr>
```