Data Wrangling - Theses project

Anh Thu

10/12/2021

```
# Import the library
library(readr)
library(naniar)
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(stringr)
library(tidyr)
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
library(ggplot2)
library(plotly)
```

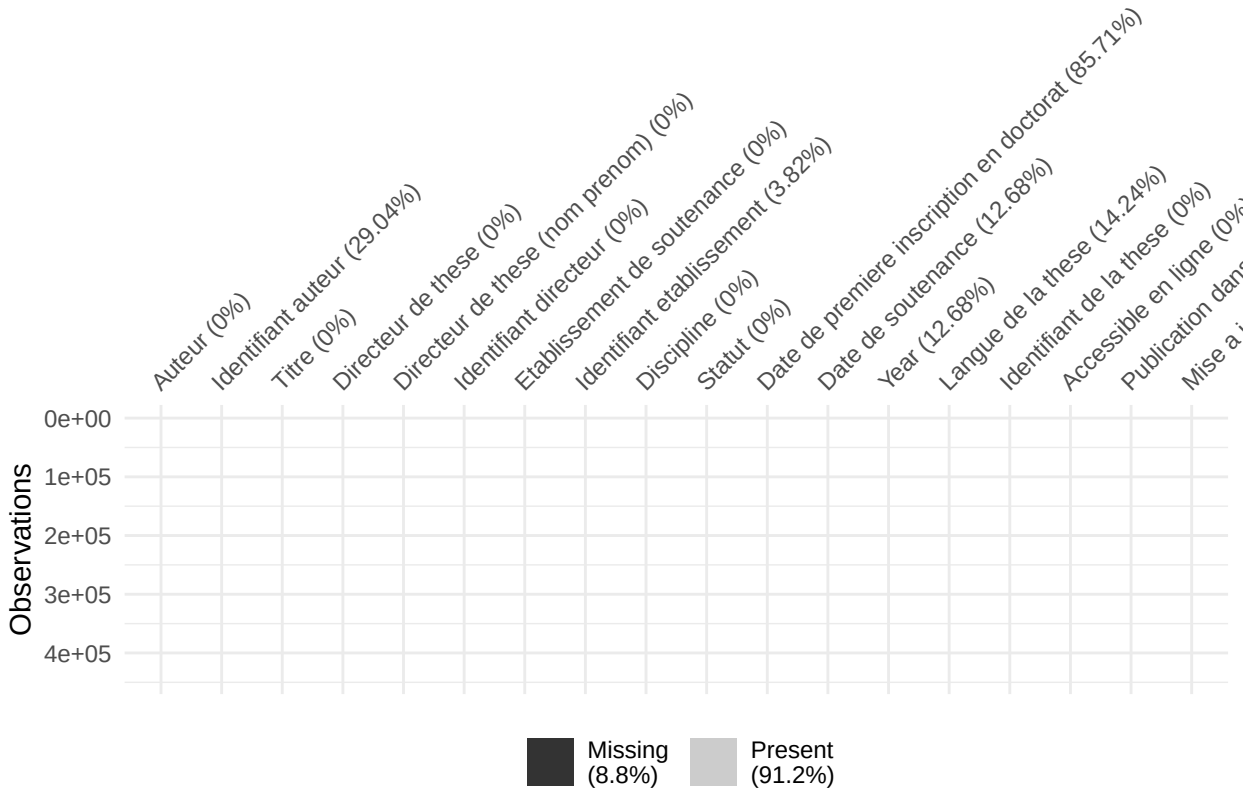
```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following objects are masked from 'package:plyr':
##
##     arrange, mutate, rename, summarise

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
# Read the csv file
df=read_csv("theses_v2.csv")
```

```
## Rows: 447644 Columns: 18

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): Auteur, Identifiant auteur, Titre, Directeur de these, Directeur d...
## dbl  (1): Year

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Missing Data

```
#Plot the missing data graph
vis_miss(df, warn_large_data = FALSE)
```



Create the n.pages variable under condition & complete missing values using an imputation technique.

```
x <- seq(1, as.integer(0.8 * nrow(df)))
y <- rnorm(x, mean = 200, sd = 50)
missing = nrow(df) - as.integer( 0.8 * nrow(df))
na_col <- rep(NA, missing)
set.seed(100)
n.pages = sample(c(as.integer(y), na_col))
df$n.pages <- n.pages
head(df$n.pages, 10)
```

```
##  [1] 255 204 292 169 200 203  NA 192  NA 169
```

Common issues

##The proportion of defences at the first of january evolve over the years

```
# Load the Date de soutenance column
str(df$"Date de soutenance")
```

```
##  chr [1:447644] NA NA "01-01-93" NA NA "24-11-08" "01-07-05" "08-12-09" ...
```

```
dt <- df$"Date de soutenance"
head(dt, n=10)
```

```
##  [1] NA          NA          "01-01-93" NA          NA          "24-11-08"
##  [7] "01-07-05" "08-12-09" "10-01-13" "24-06-11"
```

```r
# Convert to class "Date" representing calendar dates
dt <- as.Date(dt, "%d-%m-%y")
```

```r
# Create data framee df_date
df_date <- data.frame(dt)
df_date <- na.omit(df_date)
head(df_date)
```

```
##            dt
## 3  1993-01-01
## 6  2008-11-24
## 7  2005-07-01
## 8  2009-12-08
## 9  2013-01-10
## 10 2011-06-24
```

```r
# Parse and manipulate dates into different column (year, month, day)
df_date <- df_date %>% dplyr::mutate(year = lubridate::year(dt), month = lubridate::month(dt), day = lu
head(df_date)
```

```
##            dt year month day
## 3  1993-01-01 1993     1   1
## 6  2008-11-24 2008    11  24
## 7  2005-07-01 2005     7   1
## 8  2009-12-08 2009    12   8
## 9  2013-01-10 2013     1  10
## 10 2011-06-24 2011     6  24
```

```r
# Create data frame newyear
newyear <- df_date %>% filter(month == 1)
newyear<- newyear %>% filter(day == 1)
head(newyear)
```

```
##           dt year month day
## 1 1993-01-01 1993     1   1
## 2 2015-01-01 2015     1   1
## 3 2015-01-01 2015     1   1
## 4 2012-01-01 2012     1   1
## 5 2014-01-01 2014     1   1
## 6 2012-01-01 2012     1   1
```

```r
#Group and count by year in df_newyear
df_newyear <- newyear %>% select(year) %>% group_by(year) %>% count()
head(df_newyear)
```

```
##   year freq
## 1 1971    1
## 2 1972    1
## 3 1973    1
## 4 1976    1
## 5 1979    1
## 6 1980    1
```

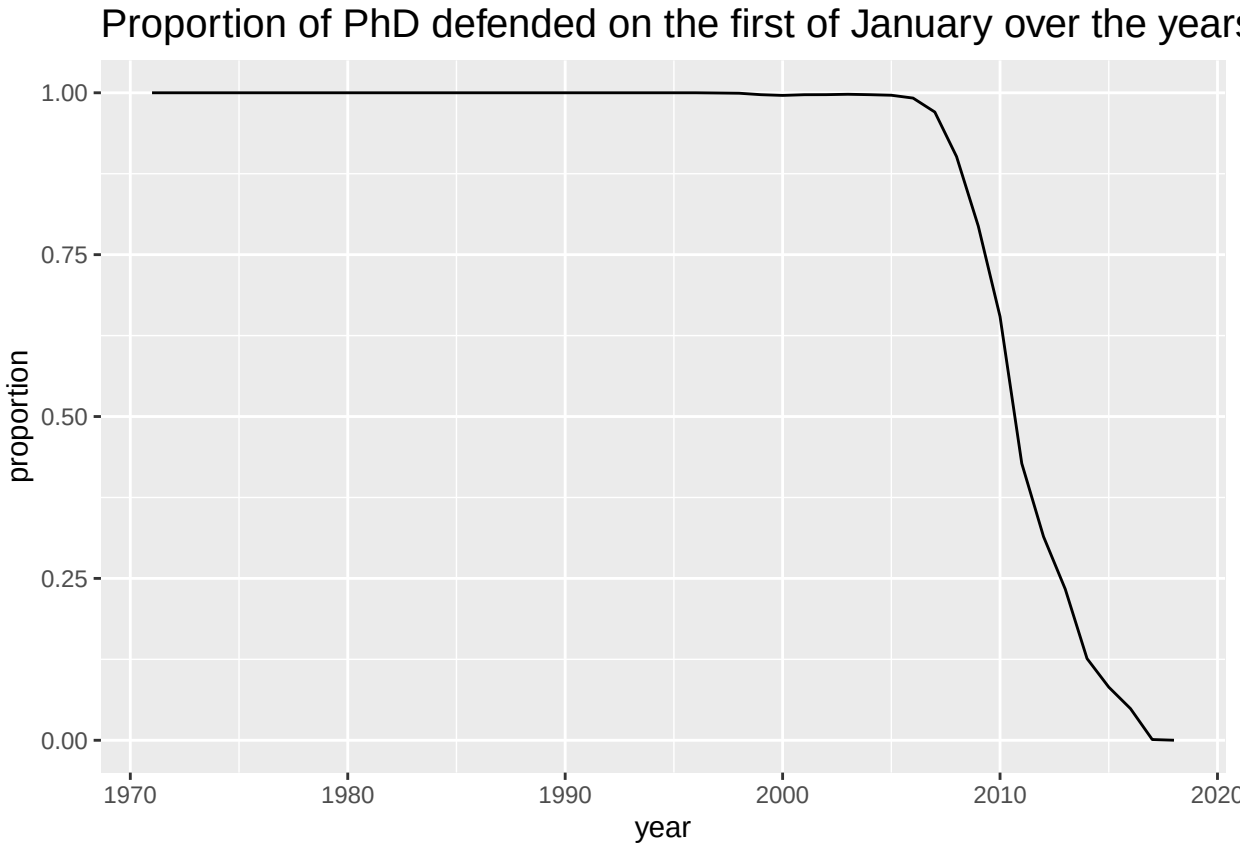
```r
#Group and count by year in df_date_all
df_date_all <-df_date %>% select(year) %>% group_by(year) %>% count()
head(df_date_all)
```

```
##   year freq
## 1 1971    1
## 2 1972    1
## 3 1973    1
## 4 1976    1
## 5 1979    1
## 6 1980    1
```

```r
# Mutating joins add columns from from df_newyear to df_date_all
df_proportion <- inner_join(df_newyear, df_date_all, by ="year")
df_proportion$proportion <- df_proportion$freq.x/df_proportion$freq.y
head(df_proportion)
```

```
##   year freq.x freq.y proportion
## 1 1971      1      1          1
## 2 1972      1      1          1
## 3 1973      1      1          1
## 4 1976      1      1          1
## 5 1979      1      1          1
## 6 1980      1      1          1
```

```r
# Plot the proportion of defences at the first of january evolve over the years
ggplot(df_proportion, aes(x=year, y=proportion)) +
  geom_line() + ggtitle("Proportion of PhD defended on the first of January over the years")+ theme(plo
```




```
subset(df_proportion, year > 2005 & year < 2015)
```

```
##    year freq.x freq.y proportion
## 30 2006  10885  10975  0.9917995
## 31 2007  11349  11697  0.9702488
## 32 2008  10686  11854  0.9014679
## 33 2009   9554  12033  0.7939832
## 34 2010   8190  12516  0.6543624
## 35 2011   5605  13110  0.4275362
## 36 2012   4398  13985  0.3144798
## 37 2013   3237  13868  0.2334151
## 38 2014   1666  13202  0.1261930
```

```
#the proportion of defenses at the first of January started decreasing from 2006 (before that it was al
```

##Cecile Martin problems

```
# Load the Cécile Martin from Auteur column in df
Cecile <- filter(df, df$Auteur == "Cecile Martin")
Cecile
```

```
## # A tibble: 7 x 19
##   Auteur        `Identifiant auteur` Titre     `Directeur de t~ `Directeur de t~
##   <chr>         <chr>                <chr>     <chr>            <chr>
## 1 Cecile Martin 203208145            L'invent~ Laurent Jullier  Jullier Laurent
```

```
## 2 Cecile Martin 81323557            Systeme ~ JEAN LOSSOUARN   LOSSOUARN JEAN
## 3 Cecile Martin 179423568           Concurre~ Brigitte Dormont Dormont Brigitte
## 4 Cecile Martin 81323557            Modelisa~ Gerard Antonini  Antonini Gerard
## 5 Cecile Martin 81323557            Caracter~ Jean Mironneau   Mironneau Jean
## 6 Cecile Martin 81323557            Influenc~ Yves Briand      Briand Yves
## 7 Cecile Martin 182118703           Depositi~ Dominique Vauth~ Vautherin Domin~
## # ... with 14 more variables: Identifiant directeur <chr>,
## #   Etablissement de soutenance <chr>, Identifiant etablissement <chr>,
## #   Discipline <chr>, Statut <chr>,
## #   Date de premiere inscription en doctorat <chr>, Date de soutenance <chr>,
## #   Year <dbl>, Langue de la these <chr>, Identifiant de la these <chr>,
## #   Accessible en ligne <chr>, Publication dans theses.fr <chr>,
## #   Mise a jour dans theses.fr <chr>, n.pages <int>
```

```
#There are 4 different people with the same name Cecile, 1 of them has 4 theses
```

Supervisor's ID

```
#Create supervisor id data frame with the length column
supervisor_id <- df$`Identifiant directeur`
df_sup <- data.frame(supervisor_id)
df_sup <- na.omit(df_sup)
df_sup$length <- nchar(df_sup$supervisor_id)
head(df_sup)
```

```
##   supervisor_id length
## 1      29561248      8
## 2   715,441,511     11
## 3      57030758      8
## 4            na      2
## 5            na      2
## 6      26941848      8
```

```
df_sup2 <- df_sup %>% select(length) %>% group_by(length) %>% count()
head(df_sup2)
```

```
##   length   freq
## 1      1   4587
## 2      2  49309
## 3      8 255680
## 4      9  78960
## 5     11  59108
```

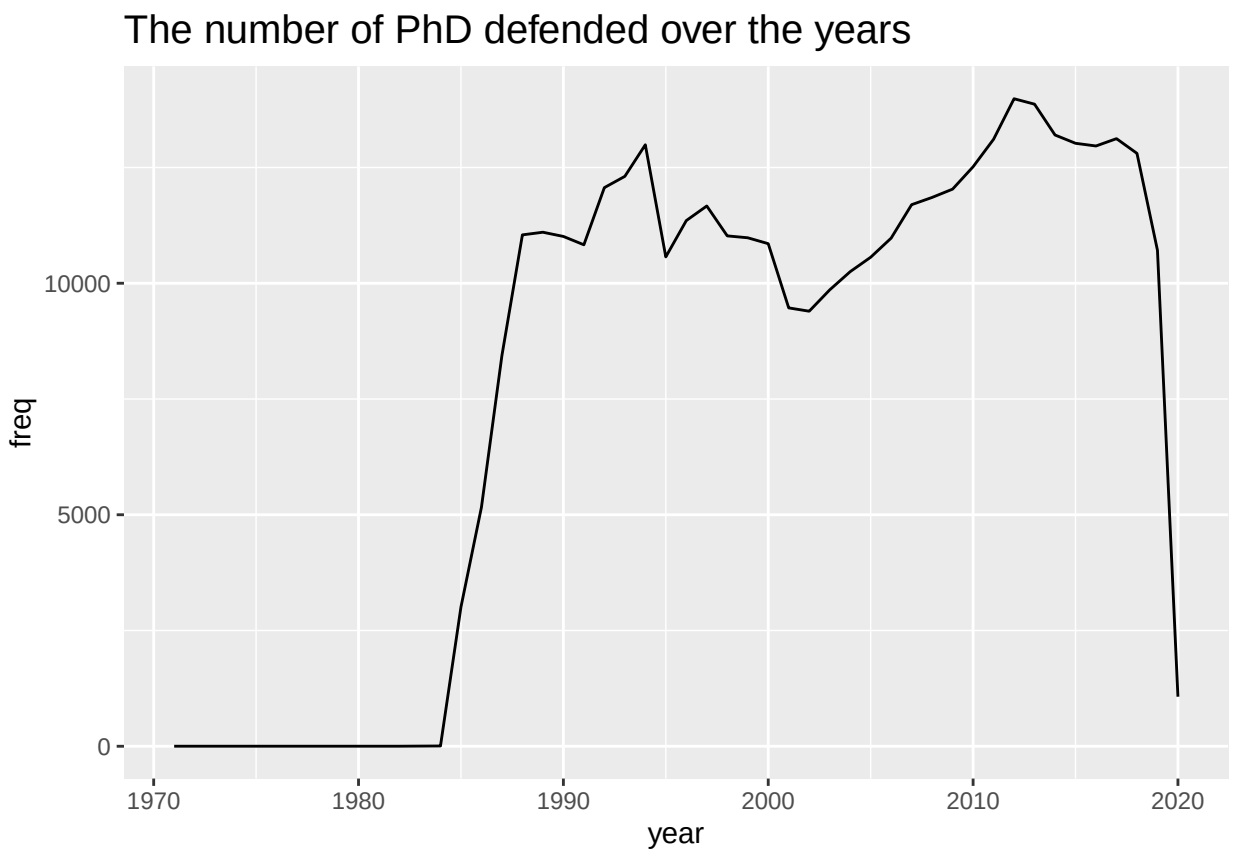
##The number of PhD defended in 2019 and 2020

```
head(df_date_all)
```

```
##   year freq
## 1 1971    1
```

```
## 2 1972    1
## 3 1973    1
## 4 1976    1
## 5 1979    1
## 6 1980    1
```

```
#Plot
ggplot(df_date_all, aes(x=year, y=freq)) +
  geom_line() + ggtitle("The number of PhD defended over the years") + theme(plot.title = element_text(
```




```
#There is a sudden drop in the number of PhD defended in 2019 and 2020.
subset(df_date_all, year > 2015 & year < 2021)
```

```
##    year  freq
## 40 2016 12965
## 41 2017 13123
## 42 2018 12805
## 43 2019 10712
## 44 2020  1070
```

#Outliers

##Supervisor

```r
#Group by same name of supervisor and same Id
supervisor <- df %>% group_by(`Directeur de these`,`Identifiant directeur`) %>% summarise(n=n()) %>% ar
```

```
## `summarise()` has grouped output by 'Directeur de these'. You can override using the `.groups` argume
```

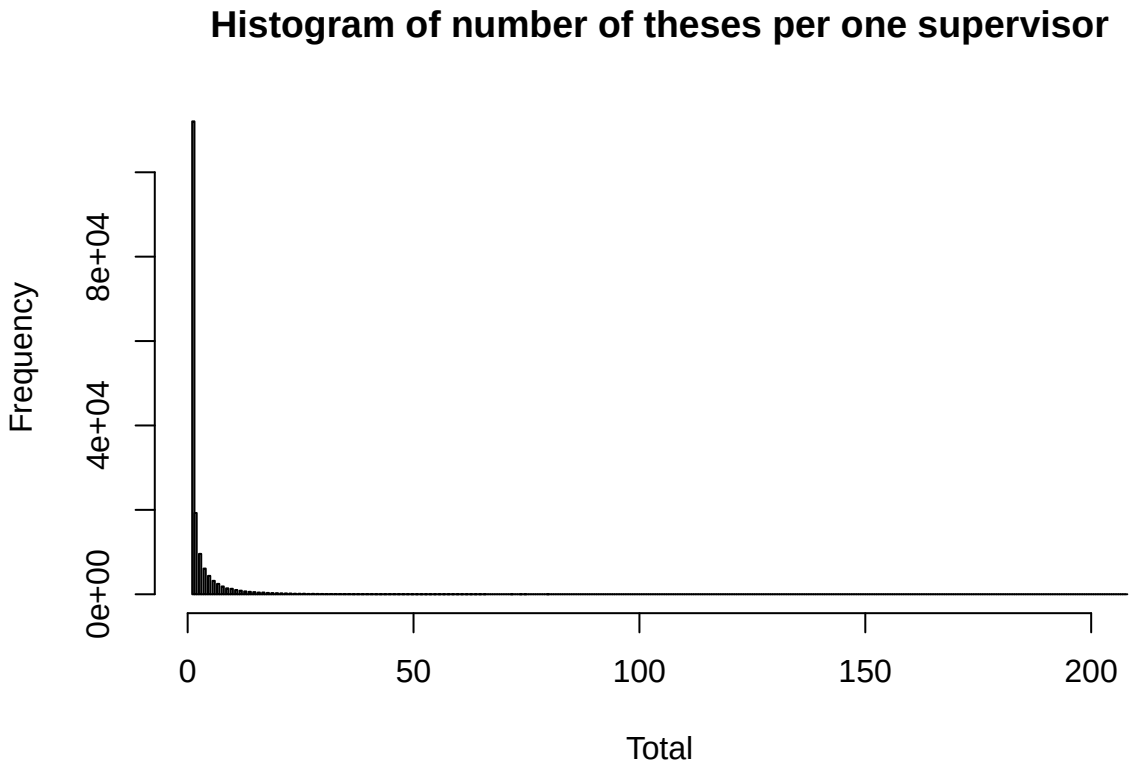
```r
supervisor <- supervisor[-1,]
head(supervisor, n=10)
```

```
## # A tibble: 10 x 3
## # Groups:   Directeur de these [10]
##    `Directeur de these`   `Identifiant directeur`     n
##    <chr>                  <chr>                   <int>
##  1 Jean-Michel Scherrmann 59375140                  208
##  2 Francois-Paul Blanc    26730774                  201
##  3 Pierre Brunel          26756625                  193
##  4 Philippe Delebecque    29561248                  178
##  5 Michel Bertucat        98531891                  173
##  6 Guy Pujolle            27084868                  172
##  7 Bernard Teyssie        27158578                  146
##  8 Bruno Foucart          26870177                  132
##  9 Henry de Lumley        26997894                  132
## 10 Jean-Claude Chaumeil   58552499                  131
```

```r
summary(supervisor$n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   2.644   2.000 208.000
```

```r
hist(supervisor$n,
  xlab = "Total",
  main = "Histogram of number of theses per one supervisor",
  breaks = sqrt(nrow(supervisor))
)
```



###Grubbs's test

```
library(outliers)
test <- grubbs.test(supervisor$n)
test
```

```
##
##  Grubbs test for one outlier
##
## data:  supervisor$n
## G = 43.42271, U = 0.98884, p-value < 2.2e-16
## alternative hypothesis: highest value 208 is an outlier
```

```
test <- grubbs.test(supervisor$n, opposite = TRUE)
test
```

```
##
##  Grubbs test for one outlier
##
## data:  supervisor$n
## G = 0.34767, U = 1.00000, p-value = 1
## alternative hypothesis: lowest value 1 is an outlier
```

```
#Check for "Jean-Michel Scherrmann"
Scherrmann <- filter(df, df$`Directeur de these` == "Jean-Michel Scherrmann")
head(Scherrmann)
```

```
## # A tibble: 6 x 19
```

```
##   Auteur                ‘Identifiant au~ Titre ‘Directeur de t~ ‘Directeur de t~
##   <chr>                 <chr>            <chr> <chr>            <chr>
## 1 Ramzi Shawahna        158089014        Expr~ Jean-Michel Sch~ Scherrmann Jean~
## 2 Leonor Vignol         <NA>             Infl~ Jean-Michel Sch~ Scherrmann Jean~
## 3 Anne J. Moulin Paccaly 97663662        Appr~ Jean-Michel Sch~ Scherrmann Jean~
## 4 Sandrine Dauchy       87464918         Expr~ Jean-Michel Sch~ Scherrmann Jean~
## 5 Severine Piot         <NA>             Eval~ Jean-Michel Sch~ Scherrmann Jean~
## 6 Sandrine Brami        <NA>             Les ~ Jean-Michel Sch~ Scherrmann Jean~
## # ... with 14 more variables: Identifiant directeur <chr>,
## #   Etablissement de soutenance <chr>, Identifiant etablissement <chr>,
## #   Discipline <chr>, Statut <chr>,
## #   Date de premiere inscription en doctorat <chr>, Date de soutenance <chr>,
## #   Year <dbl>, Langue de la these <chr>, Identifiant de la these <chr>,
## #   Accessible en ligne <chr>, Publication dans theses.fr <chr>,
## #   Mise a jour dans theses.fr <chr>, n.pages <int>
```

##Author

```
#Group by same name of Auteur and same Iduteur-Auteur
Author <- df %>% group_by(`Auteur`,`Identifiant auteur`) %>% summarise(n=n()) %>% arrange(desc(n))
```

```
## ‘summarise()‘ has grouped output by ’Auteur’. You can override using the ‘.groups‘ argument.
```

```
df_author <- data.frame(Author)
df_author <- df_author %>% drop_na()
head(df_author, n=10)
```

```
##                Auteur Identifiant.auteur n
## 1    Catherine Leport           69413916 7
## 2      Philippe Blanc           85924660 6
## 3      Thierry Martin           60151013 6
## 4     Beatrice Durand           56833776 5
## 5        Eric Renault           34296565 5
## 6     Nathalie Martin           27013340 5
## 7        Pascal Andre           55750931 5
## 8      Patrick Martin           78079365 5
## 9      Philippe Andre           61648493 5
## 10 Philippe Chevalier           66761999 5
```

Preliminary Results

Languages

```
languages_date <- df[,c("Date de soutenance","Langue de la these")]
languages_date$"Date de soutenance" <- as.Date(languages_date$"Date de soutenance", "%d-%m-%y")
head(languages_date)
```

```
## # A tibble: 6 x 2
##   ‘Date de soutenance‘ ‘Langue de la these‘
```

```
##   <date>               <chr>
## 1 NA                   <NA>
## 2 NA                   <NA>
## 3 1993-01-01           fr
## 4 NA                   <NA>
## 5 NA                   <NA>
## 6 2008-11-24           <NA>
```

```
#Create data frame df_languages_date have 2 columns Date & Language
df_languages_date <- data.frame(languages_date)
df_languages_date <- df_languages_date%>% drop_na()
df_languages_date <- data.frame(df_languages_date)
colnames(df_languages_date) <- c("Date", "Language")
head(df_languages_date)
```

```
##         Date Language
## 1 1993-01-01       fr
## 2 2015-01-01       fr
## 3 2015-01-01       fr
## 4 2013-12-07       fr
## 5 2013-11-25       fr
## 6 2013-11-22       fr
```

```
#Lower the character in Language col
df_languages_date$Language <- tolower(df_languages_date$Language)
head(df_languages_date)
```

```
##         Date Language
## 1 1993-01-01       fr
## 2 2015-01-01       fr
## 3 2015-01-01       fr
## 4 2013-12-07       fr
## 5 2013-11-25       fr
## 6 2013-11-22       fr
```

```
#Create new col "Type" using mutate
df_languages_date <- df_languages_date %>% mutate(Type = case_when(
    (Language == "en") ~ "English",
    (Language == "fr") ~ "French",
    (Language == "enfr" | Language == "fren") ~ "Bilingual",
    TRUE ~ "Other",
    ))
head(df_languages_date)
```

```
##         Date Language   Type
## 1 1993-01-01       fr French
## 2 2015-01-01       fr French
## 3 2015-01-01       fr French
## 4 2013-12-07       fr French
## 5 2013-11-25       fr French
## 6 2013-11-22       fr French
```

```
unique(df_languages_date$Type)
```

```
## [1] "French"    "English"   "Other"     "Bilingual"
```

```
df_lang_type <- df_languages_date %>% group_by(Type) %>% count()
head(df_lang_type)
```

```
#Parsing dates using lubridate and create new col "Year"
df_languages_date <- df_languages_date %>% dplyr::mutate(Year = lubridate::year(Date))
df_languages_date <- df_languages_date[order(df_languages_date$Year),]
head(df_languages_date)
```

```
##              Date Language   Type Year
## 623    1971-01-01       fr French 1971
## 619    1972-01-01       fr French 1972
## 633    1973-01-01       fr French 1973
## 624    1976-01-01       fr French 1976
## 243466 1979-01-01       fr French 1979
## 172637 1980-01-01       fr French 1980
```

```
#Check unique
unique(df_languages_date$Year)
```

```
##  [1] 1971 1972 1973 1976 1979 1980 1982 1984 1985 1986 1987 1988 1989 1990 1991
## [16] 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006
## [31] 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
```

```
#Sum by year and type of the languages
df_la_type_year <- df_languages_date %>% select(Year, Type) %>% group_by(Year, Type) %>% count()
colnames(df_la_type_year) <- c("Year", "Type", "Sum")
head(df_la_type_year)
```

```
##   Year   Type Sum
## 1 1971 French   1
## 2 1972 French   1
## 3 1973 French   1
## 4 1976 French   1
## 5 1979 French   1
## 6 1980 French   1
```

```
#Sum by year
df_year <- df_languages_date %>% select(Year) %>% group_by(Year) %>% count()
colnames(df_year) <- c("Year", "Sum_Year")
head(df_year)
```

```
##   Year Sum_Year
## 1 1971        1
## 2 1972        1
## 3 1973        1
## 4 1976        1
## 5 1979        1
## 6 1980        1
```

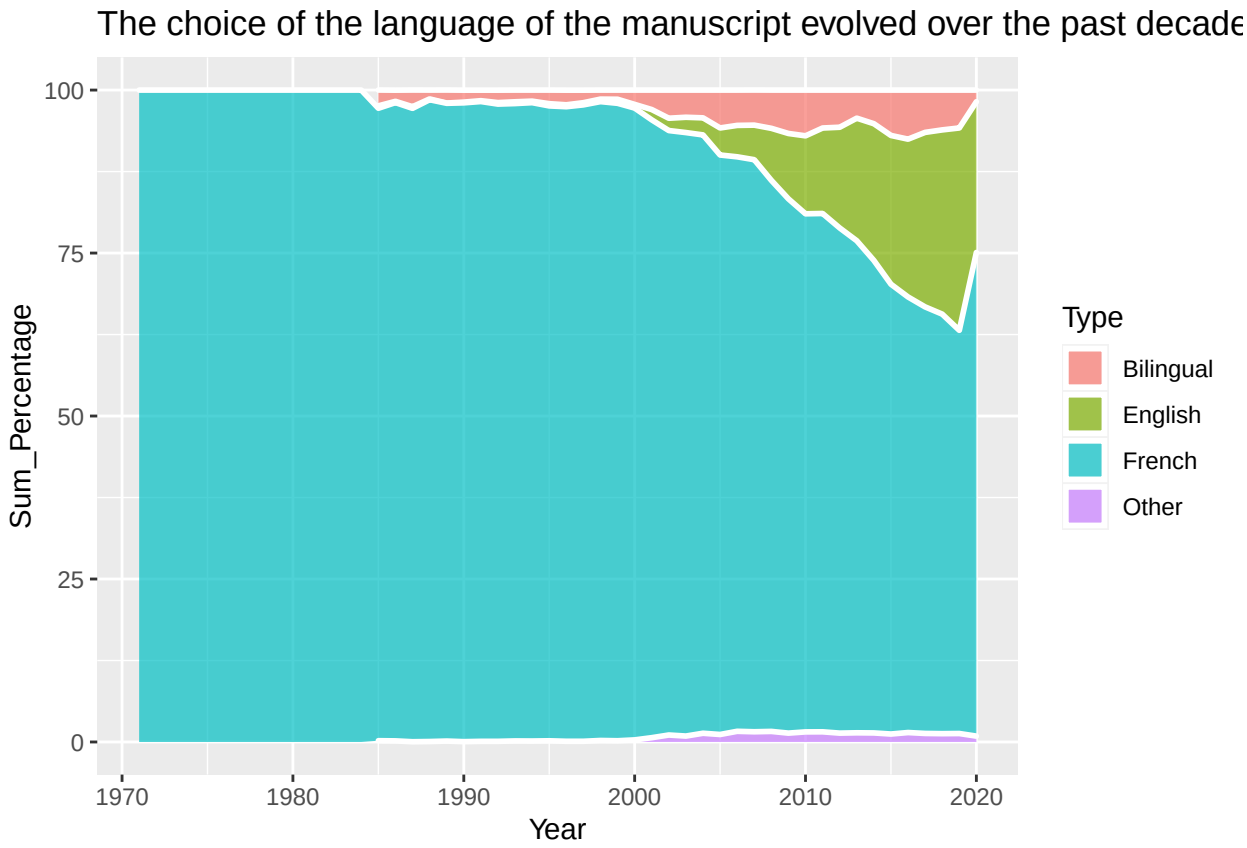
```r
#Merge 2 df
full_lang_type <- full_join(df_la_type_year, df_year, by = 'Year')
head(full_lang_type)
```

```
##   Year   Type Sum Sum_Year
## 1 1971 French   1        1
## 2 1972 French   1        1
## 3 1973 French   1        1
## 4 1976 French   1        1
## 5 1979 French   1        1
## 6 1980 French   1        1
```

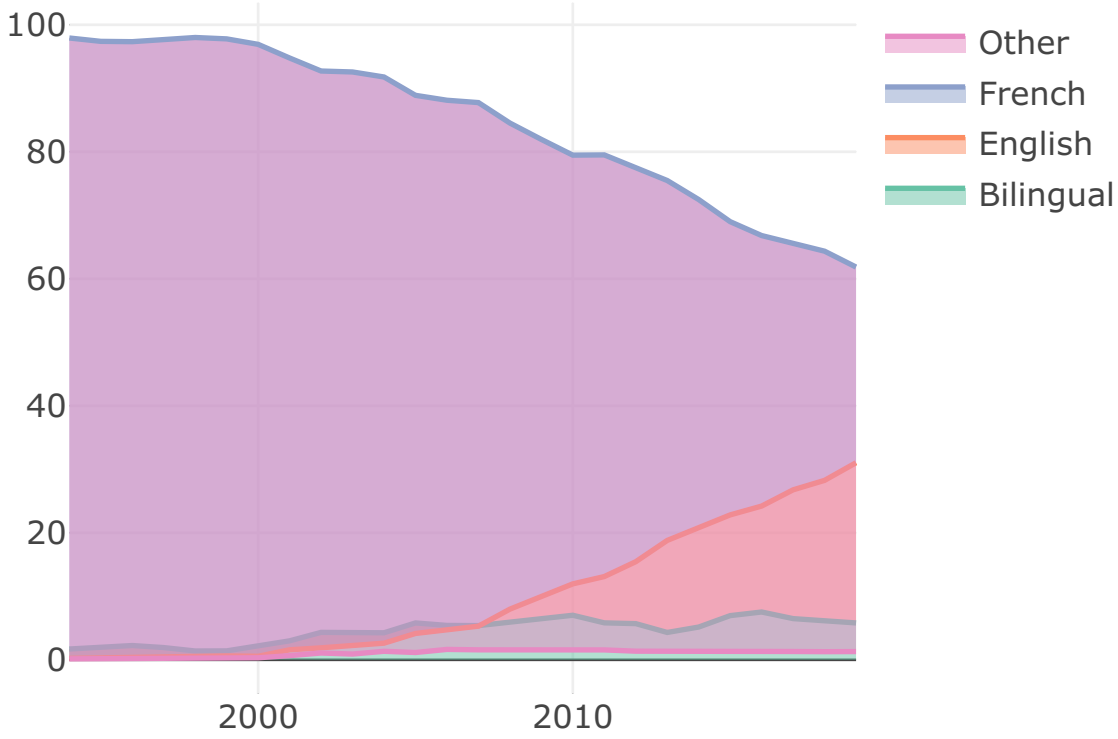
```r
#Calculate percentage of sum
full_lang_type$Sum_Percentage <- round((full_lang_type$Sum / full_lang_type$Sum_Year) * 100, 2)
head(full_lang_type)
```

```
##   Year   Type Sum Sum_Year Sum_Percentage
## 1 1971 French   1        1            100
## 2 1972 French   1        1            100
## 3 1973 French   1        1            100
## 4 1976 French   1        1            100
## 5 1979 French   1        1            100
## 6 1980 French   1        1            100
```

```r
# Plot
ggplot(full_lang_type, aes(x=Year, y=Sum_Percentage, fill=Type)) +
    geom_area(alpha=0.7 , size=1, colour="white") +
  ggtitle("The choice of the language of the manuscript evolved over the past decades")
```




```
df_plotly <- full_lang_type %>% filter(Year >= 1994 & Year <2020)
plot_ly(type = 'scatter', x = df_plotly$Year, y = df_plotly$Sum_Percentage, color = df_plotly$Type,
            mode = 'lines', fill = 'tonexty')
```



```
colnames(df)
```

```
##  [1] "Auteur"
##  [2] "Identifiant auteur"
##  [3] "Titre"
##  [4] "Directeur de these"
##  [5] "Directeur de these (nom prenom)"
##  [6] "Identifiant directeur"
##  [7] "Etablissement de soutenance"
##  [8] "Identifiant etablissement"
##  [9] "Discipline"
## [10] "Statut"
## [11] "Date de premiere inscription en doctorat"
## [12] "Date de soutenance"
## [13] "Year"
## [14] "Langue de la these"
## [15] "Identifiant de la these"
## [16] "Accessible en ligne"
## [17] "Publication dans theses.fr"
## [18] "Mise a jour dans theses.fr"
## [19] "n.pages"
```

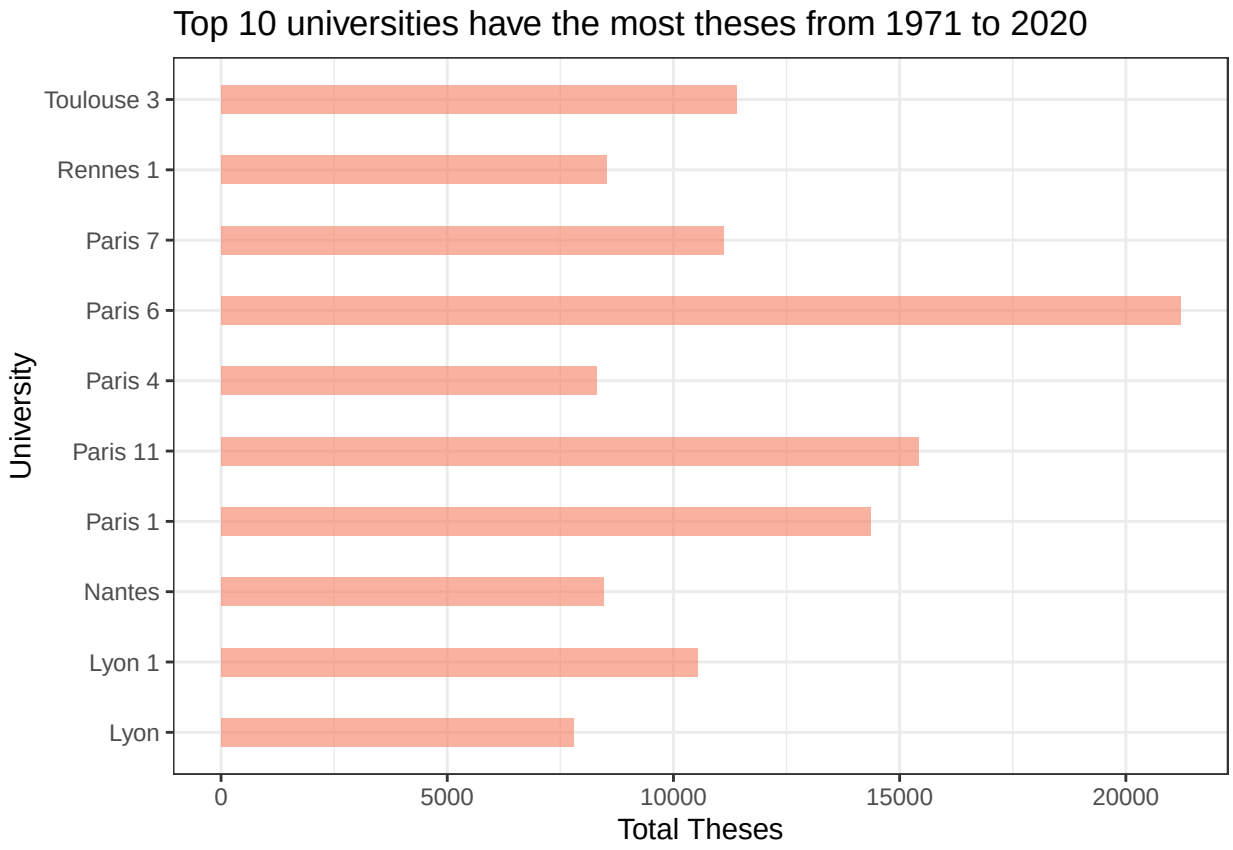
```
University <- df %>% group_by(`Etablissement de soutenance`) %>% summarise(n=n()) %>% arrange(desc(n))
uni <-head(University, n=10)
uni
```

```
## # A tibble: 10 x 2
##    `Etablissement de soutenance`     n
##    <chr>                         <int>
##  1 Paris 6                       21201
##  2 Paris 11                      15429
##  3 Paris 1                       14347
##  4 Toulouse 3                    11385
##  5 Paris 7                       11101
##  6 Lyon 1                        10522
##  7 Rennes 1                       8524
##  8 Nantes                         8455
##  9 Paris 4                        8303
## 10 Lyon                           7783
```

```
# load the library
library(forcats)

# Reorder following the value of another column:
uni %>%
  ggplot( aes(x=uni$`Etablissement de soutenance`, y=uni$n)) +
    geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
    coord_flip() +
    xlab("University") +
    ylab("Total Theses") +
    ggtitle("Top 10 universities have the most theses from 1971 to 2020") +
    theme_bw()
```

```
## Warning: Use of ‘uni$‘Etablissement de soutenance‘‘ is discouraged. Use
## ‘Etablissement de soutenance‘ instead.
```

```
## Warning: Use of ‘uni$n‘ is discouraged. Use ‘n‘ instead.
```

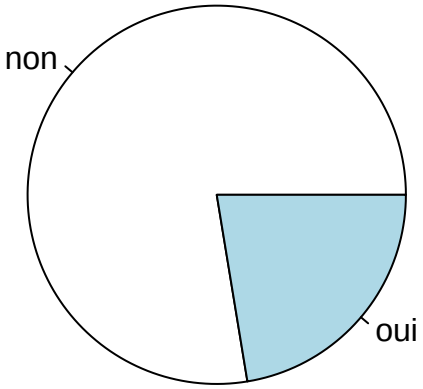



```
Public <- df %>% group_by(`Accessible en ligne`) %>% summarise(n=n()) %>% arrange(desc(n))
Public
```

```
## # A tibble: 2 x 2
##   ‘Accessible en ligne‘      n
##   <chr>                  <int>
## 1 non                   347341
## 2 oui                   100303
```

```
labels = c( "no","yes")
pie(Public$n,Public$`Accessible en ligne`, main="The protortion of theses accessable online")
```

The protortion of theses accessable online



Theses project - python

October 21, 2021

[1]:
```
#import library
import pandas as pd
import numpy as np
import missingno as msno
import gender_guesser.detector as gender
gen = gender.Detector()
import plotly.graph_objects as go
import plotly.express as px
```

[2]:
```
#Load the csv file
df = pd.read_csv("theses_v2.csv", low_memory=False)
df.head(3)
```

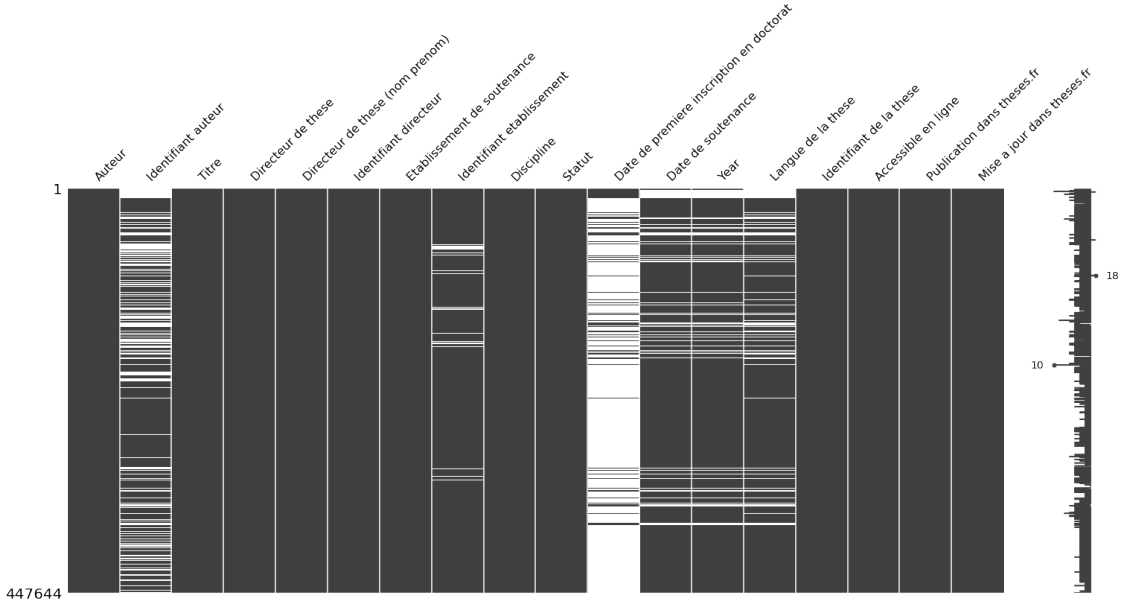
[2]:

	Auteur	Identifiant auteur	Titre	Directeur de these	Directeur de these (nom prenom)	Identifiant directeur	Etablissement de soutenance
0	Saeed Al marri	NaN	Le credit documentaire et l'onopposabilite des…	Philippe Delebecque	Delebecque Philippe	29561248	Paris 1
1	Andrea Ramazzotti	174423705	Application de la PGD a la resolution de probl…	Jean-Claude Grandidier,Marianne Beringhier	Grandidier Jean-Claude,Beringhier Marianne	715,441,511	Chasseneuil-du-Poitou, Ecole nationale superie…
2	OLIVIER BODENREIDER	NaN	Conception d'un outil informatique d'etude des…	Francois Kohler	Kohler Francois	57030758	Nancy 1

```
  Identifiant etablissement  \
0                  27361802
1                  28024400
2                       NaN

                                      Discipline    Statut  \
0                                     Driot prive   enCours
1  Mecanique des solides, des materiaux, des stru…   enCours
2                                        Medecine  soutenue

  Date de premiere inscription en doctorat Date de soutenance    Year  \
0                                 30-09-11                NaN     NaN
1                                 01-10-12                NaN     NaN
2                                      NaN           01-01-93  1993.0

  Langue de la these Identifiant de la these Accessible en ligne  \
0                NaN                  s69480                 non
1                NaN                  s98826                 non
2                 fr           1993NAN19006                 non

  Publication dans theses.fr Mise a jour dans theses.fr
0                   26-01-12                   26-01-12
1                   22-11-13                   22-11-13
2                   24-05-13                   17-11-12
```

[3]:
```
msno.matrix(df)
```

[3]: <AxesSubplot:>



[4]:
```
#Chose df from 2010
df.dropna(subset=['Date de soutenance'], inplace=True)
df['Date de soutenance'] = pd.DatetimeIndex(df['Date de soutenance'])
df = df[df['Date de soutenance'].apply(lambda x: np.logical_and(x.year > 2009,␣
 ↪np.logical_or(x.day != 1, x.month != 1)))]
```

[5]:
```
df.head(4)
```

[5]:

```
                          Auteur Identifiant auteur  \
8   Jennifer Guiraud (McKELLIPS)                NaN
9     Nathalie Warcholak (David)                NaN
10  Scheherazade Pinilla canadas                NaN
15                Elodie Demaret                NaN

                                                 Titre      Directeur de these  \
8   L'autobiographie sans frontieres : espace et d…  Anne-Emmanuelle Berger
9               Interoperabilite et droits du marche.     Jean-Pierre Clavier
10  Les cites reapparaissantes: L'heroisme du gran…        Patrice Vermeren
15  La mediation comme facteur de maitrise intelle…       Emile-Henri Riard

   Directeur de these (nom prenom) Identifiant directeur  \
8           Berger Anne-Emmanuelle               32574088
9              Clavier Jean-Pierre               35557060
10                Vermeren Patrice               28251873
15               Riard Emile-Henri              137391919

   Etablissement de soutenance Identifiant etablissement  \
8                      Paris 8                  26403552
9                       Nantes                  26403447
10                     Paris 8                  26403552
15                      Amiens                  26403714

                                        Discipline   Statut  \
8                                 Etudes de genre  enCours
9                                    Droit prive  enCours
10  Philosophie (metaphysique, epistemologie, esth…  enCours
15                                   Psychologie  enCours

   Date de premiere inscription en doctorat Date de soutenance    Year  \
8                                  01-11-03         2013-10-01  2013.0
9                                  01-12-02         2011-06-24  2011.0
10                                 01-03-03         2010-11-26  2010.0
15                                 01-11-03         2011-06-10  2011.0

   Langue de la these Identifiant de la these Accessible en ligne  \
```

```
8                NaN                 s11354                 non
9                NaN                  s9544                 non
10               NaN                 s11451                 non
15               NaN                  s9649                 non

   Publication dans theses.fr Mise a jour dans theses.fr
8                    26-09-11                    04-04-16
9                    26-09-11                    05-04-16
10                   26-09-11                    02-04-12
15                   26-09-11                    06-02-12
```

[6]:
```
#Create col "Month" & "Year"
df['Month'] = df['Date de soutenance'].apply(lambda x: x.month)
df['Year'] = df['Date de soutenance'].apply(lambda x: x.year)
df.head(4)
```

[6]:
```
                          Auteur Identifiant auteur  \
8   Jennifer Guiraud (McKELLIPS)                NaN
9     Nathalie Warcholak (David)                NaN
10  Scheherazade Pinilla canadas                NaN
15                Elodie Demaret                NaN

                                                 Titre      Directeur de these  \
8   L'autobiographie sans frontieres : espace et d…  Anne-Emmanuelle Berger
9                Interoperabilite et droits du marche.     Jean-Pierre Clavier
10  Les cites reapparaissantes: L'heroisme du gran…        Patrice Vermeren
15  La mediation comme facteur de maitrise intelle…       Emile-Henri Riard

   Directeur de these (nom prenom) Identifiant directeur  \
8           Berger Anne-Emmanuelle               32574088
9              Clavier Jean-Pierre               35557060
10                Vermeren Patrice               28251873
15               Riard Emile-Henri              137391919

   Etablissement de soutenance Identifiant etablissement  \
8                      Paris 8                  26403552
9                       Nantes                  26403447
10                     Paris 8                  26403552
15                      Amiens                  26403714

                                        Discipline   Statut  \
8                                 Etudes de genre  enCours
9                                     Droit prive  enCours
10  Philosophie (metaphysique, epistemologie, esth…  enCours
15                                    Psychologie  enCours

   Date de premiere inscription en doctorat Date de soutenance  Year  \
```

```
8                                   01-11-03         2013-10-01  2013
9                                   01-12-02         2011-06-24  2011
10                                  01-03-03         2010-11-26  2010
15                                  01-11-03         2011-06-10  2011

   Langue de la these Identifiant de la these Accessible en ligne  \
8                 NaN                  s11354                 non
9                 NaN                   s9544                 non
10                NaN                  s11451                 non
15                NaN                   s9649                 non

   Publication dans theses.fr Mise a jour dans theses.fr  Month
8                    26-09-11                   04-04-16     10
9                    26-09-11                   05-04-16      6
10                   26-09-11                   02-04-12     11
15                   26-09-11                   06-02-12      6
```

[7]:
```
years = df.groupby('Year').count().reset_index().reindex(['Year', 'Titre'],␣
↪axis=1).set_index('Year')
years.head(4)
```

[7]:
```
      Titre
Year
2010   4326
2011   7505
2012   9587
2013  10631
```

[8]:
```
#Create df_months from df and calculate the percentage
df_months = df.groupby(['Year', 'Month']).count().reset_index().
↪reindex(['Year', 'Month', 'Titre'], axis=1)
df_months['nb_Year'] = df_months['Year'].apply(lambda x: years.loc[x])
df_months['Percentage'] = df_months['Titre'] / df_months['nb_Year'] * 100
df_months['Time'] = pd.to_datetime(df_months[['Year', 'Month']].assign(day=1))
df_months.head(4)
```

[8]:
```
   Year  Month  Titre  nb_Year  Percentage       Time
0  2010      1    268     4326    6.195099 2010-01-01
1  2010      2    176     4326    4.068423 2010-02-01
2  2010      3    287     4326    6.634304 2010-03-01
3  2010      4    205     4326    4.738789 2010-04-01
```

[9]:
```
#Calculate the mean value
df_test=pd.DataFrame(df_months.groupby(['Month'])['Percentage'].mean())
df_test
```

[9]:

Month	Percentage
1	7.832324
2	5.455898
3	6.341093
4	5.071927
5	6.620925
6	9.864329
7	4.563513
8	3.515359
9	10.941472
10	9.794319
11	14.314484
12	15.684357

[10]:
```
#Calculate the std value
df_test2 = pd.DataFrame(df_months.groupby(['Month'])['Percentage'].std())
df_test2
```

[10]:

Month	Percentage
1	6.391824
2	3.050201
3	0.671306
4	0.530827
5	1.708365
6	1.683027
7	0.752937
8	0.889911
9	3.083066
10	1.802613
11	3.889338
12	4.522215

[11]:
```
df_test2.rename(columns={"Percentage":"sd"},inplace=True)
df_test2
```

[11]:

Month	sd
1	6.391824
2	3.050201
3	0.671306
4	0.530827
5	1.708365
6	1.683027
7	0.752937
8	0.889911

```
9      3.083066
10     1.802613
11     3.889338
12     4.522215
```

[12]:
```
df_thesis = pd.merge(df_test,df_test2,on="Month")
df_thesis.reset_index(inplace=True)
df_thesis
```

[12]:

	Month	Percentage	sd
0	1	7.832324	6.391824
1	2	5.455898	3.050201
2	3	6.341093	0.671306
3	4	5.071927	0.530827
4	5	6.620925	1.708365
5	6	9.864329	1.683027
6	7	4.563513	0.752937
7	8	3.515359	0.889911
8	9	10.941472	3.083066
9	10	9.794319	1.802613
10	11	14.314484	3.889338
11	12	15.684357	4.522215

[13]:
```
#Covert month to name
import datetime
def monthnum_toname(x):
    month = datetime.date(1900, x, 1).strftime('%B')
    return month
```

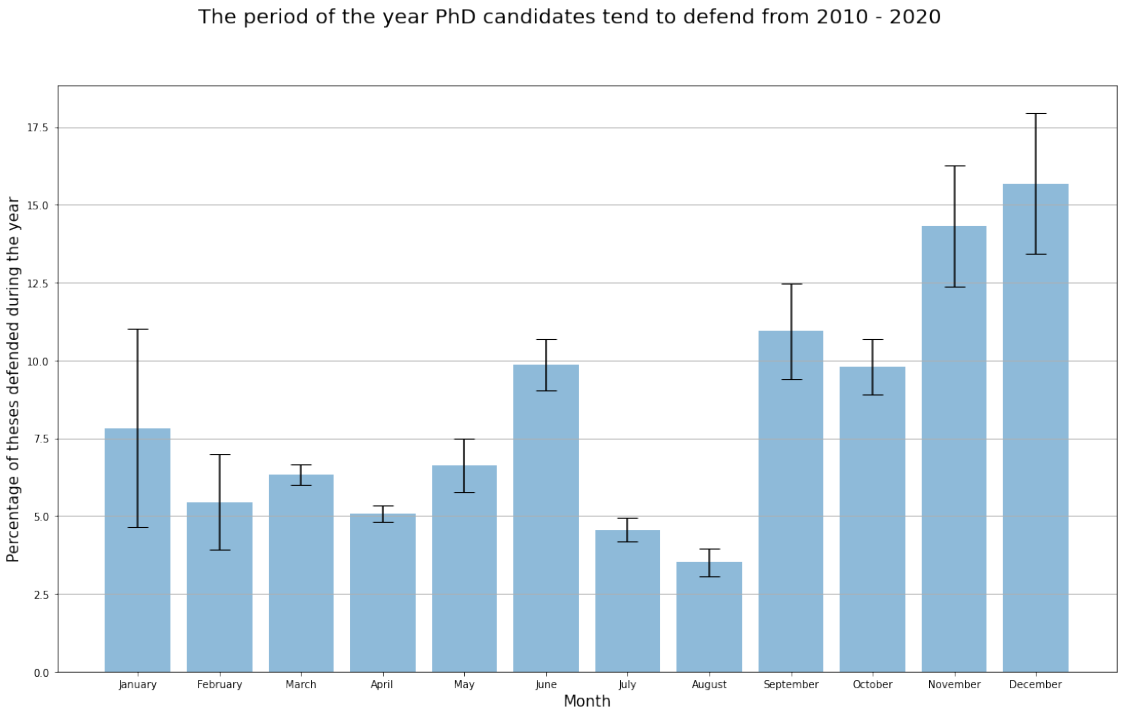
[14]:
```
df_thesis["month_name"]=df_thesis["Month"].apply(lambda x:monthnum_toname(x))
df_thesis["month_name"]
```

[14]:
```
0       January
1      February
2         March
3         April
4           May
5          June
6          July
7        August
8     September
9       October
10     November
11     December
Name: month_name, dtype: object
```

[15]:
```
df_thesis.to_csv("df_thesis.csv",)
```

[16]:
```
import matplotlib.pyplot as plt
```

[17]:
```
#Plot
fig, ax = plt.subplots()
ax.bar(df_thesis["month_name"], df_thesis["Percentage"], yerr=df_thesis["sd"]/
 ↪2, align='center', alpha=0.5, ecolor='black', capsize=10)
ax.set_ylabel('Percentage of theses defended during the year',fontsize=15)
ax.set_xlabel('Month',fontsize=15)
ax.set_xticks(df_thesis["month_name"])
ax.set_xticklabels(df_thesis["month_name"])
fig.suptitle('The period of the year PhD candidates tend to defend from 2010 -␣
 ↪2020', fontsize=20)
ax.yaxis.grid(True)
fig.set_size_inches(18.5, 10.5)
```



[18]:
```
gender = df[["Auteur","Date de soutenance"]]
```

[19]:
```
#Split the column Auteur in order to create "First name" column
gender['First_name']=gender.loc[:, ('Auteur')].str.split(expand=True)[[0]]
gender.head(4)
```

```
<ipython-input-19-39e05f11f8ea>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  gender['First_name']=gender.loc[:, ('Auteur')].str.split(expand=True)[[0]]
```

[19]:

	Auteur	Date de soutenance	First_name
8	Jennifer Guiraud (McKELLIPS)	2013-10-01	Jennifer
9	Nathalie Warcholak (David)	2011-06-24	Nathalie
10	Scheherazade Pinilla canadas	2010-11-26	Scheherazade
15	Elodie Demaret	2011-06-10	Elodie

[20]:
```
#Create function to find the gender by using gender_guesser.detector library
def get_gender(x,gen):
    return gen.get_gender(u"{}".format(x))
```

[21]:
```
#Apply function get_gender
gender["Gender"] = gender['First_name'].apply(lambda x:get_gender(x,gen))
gender.head(4)
```

```
<ipython-input-21-ce6313b0b0ce>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  gender["Gender"] = gender['First_name'].apply(lambda x:get_gender(x,gen))
```

[21]:

	Auteur	Date de soutenance	First_name	Gender
8	Jennifer Guiraud (McKELLIPS)	2013-10-01	Jennifer	female
9	Nathalie Warcholak (David)	2011-06-24	Nathalie	female
10	Scheherazade Pinilla canadas	2010-11-26	Scheherazade	unknown
15	Elodie Demaret	2011-06-10	Elodie	female

[22]:
```
#Create col "Year" in gender
gender['Year'] = pd.DatetimeIndex(gender["Date de soutenance"]).year
```

```
<ipython-input-22-9adc5d2e4d46>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  gender['Year'] = pd.DatetimeIndex(gender["Date de soutenance"]).year
```

[23]:
```
#Cleaning data
gender.dropna(subset=['Year'],how='all',inplace=True)
gender.isnull().sum()
gender.head(4)
```

```
<ipython-input-23-8a78a2a48150>:2: SettingWithCopyWarning:
```

```
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  gender.dropna(subset=['Year'],how='all',inplace=True)
```

[23]:
```
                          Auteur Date de soutenance    First_name   Gender  \
8   Jennifer Guiraud (McKELLIPS)         2013-10-01      Jennifer   female
9     Nathalie Warcholak (David)         2011-06-24      Nathalie   female
10  Scheherazade Pinilla canadas         2010-11-26  Scheherazade  unknown
15                Elodie Demaret         2011-06-10        Elodie   female

    Year
8   2013
9   2011
10  2010
15  2011
```

[24]:
```
gender_df = gender.groupby(['Gender','Year']).count().reset_index()
gender_df
```

[24]:

	Gender	Year	Auteur	Date de soutenance	First_name
0	andy	2010	87	87	87
1	andy	2011	155	155	155
2	andy	2012	217	217	217
3	andy	2013	242	242	242
4	andy	2014	279	279	279
..	…	…	…	…	…
61	unknown	2016	2327	2327	2327
62	unknown	2017	2582	2582	2582
63	unknown	2018	2400	2400	2400
64	unknown	2019	2080	2080	2079
65	unknown	2020	214	214	214

[66 rows x 5 columns]

[25]:
```
fig = px.area(gender_df,title='The evolution of gender among PhD candidates␣
 ↪over the past decades', x="Year",␣
 ↪y="Auteur",color="Gender",line_group="Gender")
fig.show()
```

[ ]:

[ ]: