Data Wrangling - Data Processing

3.1 Scrapping

> Done in Python

3.2 Missing Data

> Import packages

In [ ]:

```
library ( "readr" )
library ( "tidyverse" )
library ( "naniar" )
library ( "sjmisc" )
library ( "tidyr" )
library ( "EnvStats" )
library ( "ggplot2" )
library ( "cowplot" )
library ( "gridExtra" )
library ( "plyr" )
library ( "plotly")
library ( "viridis" )
library( "hrbrthemes" )
library ( "xtable" )
```

> Import and clean theses_v2 dataset

In [2]:

```
theses_df  =  read_csv ( "../data/theses_v2.csv" )
head ( theses_df )
```

```
Parsed with column specification:
collars (
  Author = col_character () ,
  `Author identifier` = col_double () ,
  Title = col_character () ,
  `Director of these` = col_character () ,
  ` Thesis director (name first name)` = col_character () ,
  `Manager ID` = col_character () ,
  ` Defense institution` = col_character () ,
  `Establishment identifier` = col_character () ,
  Discipline = col_character () ,
  Status = col_character () ,
  `Date of first registration in doctorate` = col_character () ,
  ` Defense date` = col_character () ,
  Year = col_double () ,
  `Language of thesis` = col_character () ,
  `Identifier of these` = col_character () ,
  `Accessible online` = col_character () ,
  `Publication in theses.fr` = col_character () ,
  `Update in theses.fr` = col_character ()
)
Warning message:
"29831 parsing failures.
 row col expected actual file
3086 Author identifier no trailing characters X '../data/theses_v2.csv'
```

```
3121 Author identifier no trailing characters X '../data/theses_v2.csv'
3131 Author identifier no trailing characters X '../data/theses_v2.csv'
3154 Author identifier no trailing characters X '../data/theses_v2.csv'
3163 Author identifier no trailing characters X '../data/theses_v2.csv'
.... ................. .................... ...... ......................
See problems (...) for more details.
"
```

Author	Author ID	Title	Supervisor	Thesis director (name first name)	Manager ID	Defense institution	Insti
Saeed al marri	N / A	Documentary credit and the enforceability of exceptions	Philippe Delebecque	Delebecque Philippe	29561248	Paris 1	273
Andrea Ramazzotti	174423705	Application of the PGD to the resolution of transient couples problems with a view to the lightening of composite structures.	Jean-Claude Grandidier, Marianne Beringhier	Grandidier Jean-Claude, Beringhier Marianne	715,441,511	Chasseneuil-du-Poitou, National Higher School of Mechanics and Aerotechnics	280
OLIVIER BODENREIDER	N / A	Design of a computer tool for the study of kinetics observed in clinical toxicology	Francois Kohler	Kohler Francois	57030758	Nancy 1	
Emmanuel Porte	N / A	Socio-history of public policies in social matters concerning students.	Gilles Pollet	Pollet Gilles	n / A	Lyon 2	0264
Arthur devriendt	N / A	INFORMATION AND COMMUNICATION TECHNOLOGIES AND NEW RURALITIES.	Gabriel Dupuy	Dupuy Gabriel	n / A	Paris 1	273

Author	Author ID	Title	Supervisor	Thesis director (name first name)	Manager ID	Defense institution	Insti
Elmantsr Briak	N / A	Forced integration of sub-Saharan Africa in the process of globalization "structuring of economies", "destructuring of states".	Edmond Jouve	Jouve Edmond	26941848	Paris 5	264

In [3]:

```
# change spaces in column names to dots
names ( theses_df ) <-  make.names ( names ( theses_df ), unique = TRUE )
names ( theses_df )
```

1. 'Author'
2. 'Author ID'
3. 'Title'
4. 'Supervisor'
5. 'Director.of.thesis..name.firstname.'
6. 'Manager ID'
7. 'Establishment.of.support'
8. 'Institution.identifier'
9. 'Disciplined'
10. 'Status'
11. 'Date.of.first.registered.in. a.doctorate'
12. 'Date.of.support'
13. 'Year'
14. 'Language.of.thesis'
15. 'Identifier.of.thesis'
16. 'Accessible.en.ligne'
17. 'Publication.dans.theses.fr'
18. 'Update.in.theses.fr'

In [4]:

```
theses_df [ theses_df  ==  "na" ]  <-  NA
```

In [5]:

```
# verify datatypes
str ( theses_df )
```

```
spec_tbl_df [447.644 x 18] (S3: spec_tbl_df / tbl_df / tbl / data.frame)
 $ Author: chr [1: 447644] "Saeed Al marri" "Andrea Ramazzotti" "OLIVIER BODENREIDE
R" "Emmanuel Porte" ...
 $ Author ID: num [1: 447644] NA 1.74e + 08 NA NA NA ...
 $ Title: chr [1: 447644] "Documentary credit and the enforceability of exceptions"
 "Application of the PGD to the resolution of transient coupled problems with a view
to the lightening of composite structures." "Design of a computer tool for the study
of kinetics observed in clinical toxicology" "Socio-history of public policies in so
cial matters concerning students." ...
 $ Director of.these: chr [1: 447644] "Philippe Delebecque" "Jean-Claude Grandidier,
Marianne Beringhier" "Francois Kohler" "Gilles Pollet" ...
 $ Director.of.thesis..name.firstname. : chr [1: 447644] "Delebecque Philippe" "Gran
didier Jean-Claude, Beringhier Marianne" "Kohler Francois" "Pollet Gilles" ...
 $ Manager ID: chr [1: 447644] "29561248" "715,441,511" "57030758" NA ...
 $ Etablissement.de.soutenance: chr [1: 447644] "Paris 1" "Chasseneuil-du-Poitou, Na
tional Superior School of Mechanics and Aerotechnics" "Nancy 1" "Lyon 2" ...
 $ Establishment.identifier: chr [1: 447644] "27361802" "28024400" NA "02640334X"
 ...
 $ Discipline: chr [1: 447644] "Driot prive" "Mechanics of solids, materials, struct
ures and surfaces" "Medicine" "Political science" ...
 $ Status: chr [1: 447644] "in progress" "in progress" "sustained" "in progress" ...
 $ Date.of.first.doctoral.rescription: chr [1: 447644] "30-09-11" "01-10-12" NA "01-
06-11" ...
 $ Date.of.support: chr [1: 447644] NA NA "01-01-93" NA ...
 $ Year: num [1: 447644] NA NA 1993 NA NA ...
 $ Langue.de.the.these: chr [1: 447644] NA NA "fr" NA ...
 $ Identifier.of.these: chr [1: 447644] "s69480" "s98826" "1993NAN19006" "s88867"
 ...
 $ Accessible.online: chr [1: 447644] "no" "no" "no" "no" ...
 $ Publication.dans.theses.fr: chr [1: 447644] "26-01-12" "22-11-13" "24-05-13" "12-
07-13" ...
 $ Mise.a.jour.dans.theses.fr: chr [1: 447644] "26-01-12" "22-11-13" "17-11-12" "12-
01-16" ...
 - attr (*, "problems") = tibble [29.831 x 5] (S3: tbl_df / tbl / data.frame)
  .. $ row: int [1: 29831] 3086 3121 3131 3154 3163 3182 3225 3251 3261 3278 ...
  .. $ col: chr [1: 29831] "Author identifier" "Author identifier" "Author identifie
r" "Author identifier" ...
  .. $ expected: chr [1: 29831] "no trailing characters" "no trailing characters" "n
o trailing characters" "no trailing characters" ...
  .. $ actual: chr [1: 29831] "X" "X" "X" "X" ...
  .. $ file: chr [1: 29831] "'../data/theses_v2.csv'" "'../data/theses_v2.csv'"
 "'../data/theses_v2.csv'" "'. ./data/theses_v2.csv '"...
 - attr (*, "spec") =
  .. collars (
  .. Author = col_character () ,
  .. `Author identifier` = col_double () ,
  .. Title = col_character () ,
  .. `Director of these` = col_character () ,
  .. `Director of these (last name first name)` = col_character () ,
  .. `Director identifier` = col_character () ,
  .. ` Defense institution` = col_character () ,
  .. `Institution identifier` = col_character () ,
  .. Discipline = col_character () ,
  .. Status = col_character () ,
  .. `Date of first registration in doctorate` = col_character () ,
  .. ` Defense date` = col_character () ,
  .. Year = col_double () ,
  .. `Language of these` = col_character () ,
  .. `Identifier of these` = col_character () ,
  .. `Available online` = col_character () ,
  .. `Publication in theses.fr` = col_character () ,
  .. `Update in theses.fr` = col_character ()
  ..)
```

In [6]:

```
# change to date format
theses_df$Date.de.premiere.inscription.en.doctorat <- as.Date(theses_df$Date.de.prem
theses_df$Date.de.soutenance <- as.Date(theses_df$Date.de.soutenance, "%d-%m-%y")
```

Visualize missing data

In [7]:

```
# check num of missing values
n_miss(theses_df)
```

787238

In [8]:

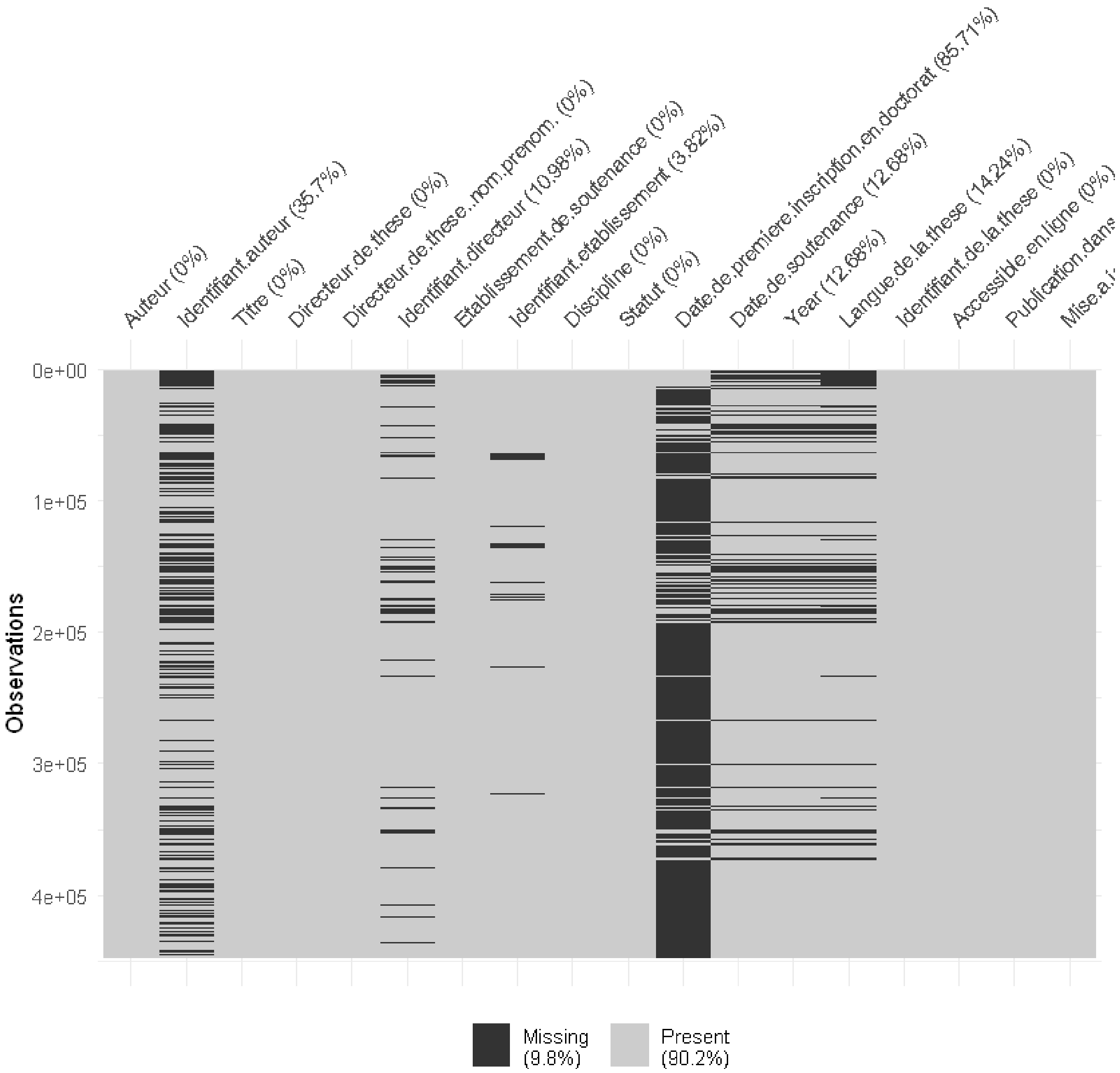
```
# get table of missing values in each column
missing_values <- theses_df %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing =is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n(), perc.missing = round((n() / 447644) * 100 ,  2 ))
    filter ( is.missing  ==  T )  %>%
    select ( - is.missing )  %>%
    arrange ( desc ( num.missing ))
# colnames (missing_values) <- c ("Column Name "," Number of Missing Values ")
xtable ( missing_values )
```

```
Warning message:
"attributes are not identical across measure variables;
they will be dropped "` summarize () `has grouped output by 'key'. You can override
 using the` .groups` argument.
```

key	num.missing	perc.missing
Doctoral.Registration Date	383668	85.71
Author ID	159820	35.70
Language.of.thesis	63765	14.24
Support date	56746	12.68
Year	56746	12.68
Manager ID	49172	10.98
Institution.identifier	17085	3.82
Update.in.theses.fr	177	0.04
Supervisor	17	0.00
Director.of.thesis..name.firstname.	17	0.00
Title	13	0.00
Disciplined	5	0.00
Support.Establishment	4	0.00
Author	3	0.00

In [9]:

```
# visualize percentage of missing data
vis_miss ( theses_df ,  warn_large_data  =  FALSE )
```



In almost all cases if the theses defense date is known the beginning date is not. Meaning that when the defense date of a theses is added the beginning date is removed.

In [14]:

```
# create n.pages for 80% of dataset with mean = 200 & sd = 50 and rest 20% as na val
x  <-  seq ( 1 ,  as.integer ( 0.8  *  nrow ( theses_df )))
y  <-  rnorm ( x ,  mean  =  200 ,  sd  =  50 )
n_missing  =  nrow ( theses_df )  -  as.integer (  0.8  *  nrow ( theses_df ))
na_col  <-  rep ( NA,  n_missing )
set.seed ( 200 )
n.pages  =  sample ( c ( as.integer ( y ),  na_col ))
theses_df $ n.pages  <-  n.pages
head ( theses_df $ n.pages ,  10 )
```

1. <NA>
2. <NA>
3. 169
4. 168
5. 277
6. 248
7. 166
8. 212

9. 162
10. 207

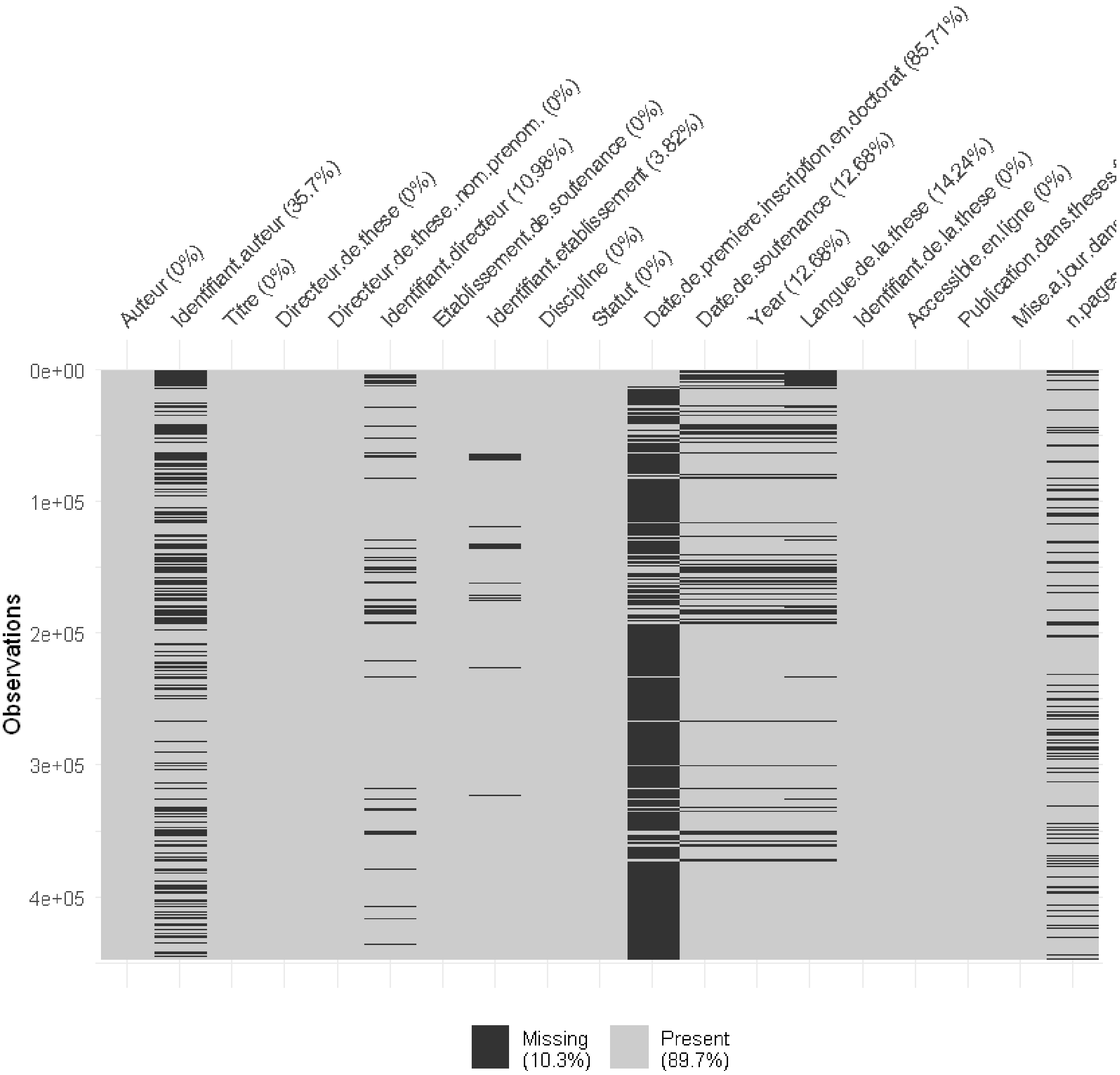
In [15]:

```
sum ( is.na ( theses_df $ n.pages ))
```

89529

In [16]:

```
# visualize percentage of missing data
vis_miss ( theses_df ,  warn_large_data  =  FALSE )
```



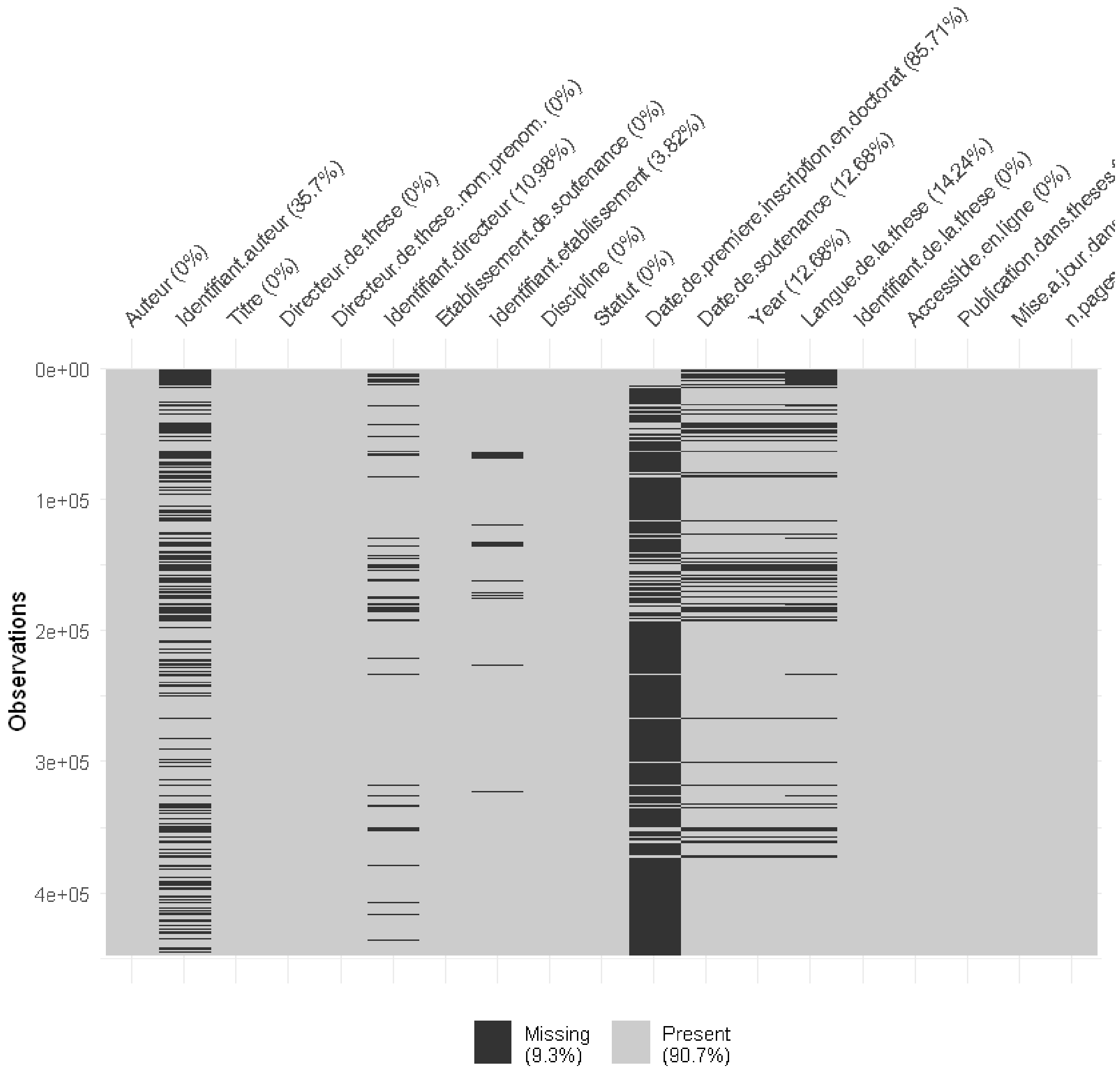
In [17]:

```
# imputate missingvaluesusing mean of n.pages
theses_df $ n.pages [ is.na ( theses_df $ n.pages )]  <-  mean ( theses_df $ n.pages
sum ( is.na ( theses_df $ n .pages ))
```

0

In [18]:

```
# visualize percentage of missing data after imputation
vis_miss ( theses_df ,  warn_large_data  =  FALSE )
```



3.3. Common Issues

> Issues in the defense data

In [19]:

```
# select defense date
defense_date <- theses_df $ Date.de.soutenance
str ( defense_date )
sum ( is.na ( defense_date ))
```

```
 Date [1: 447644], format: NA NA "1993-01-01" NA NA "2008-11-24" "2005-07-01" "2009-
12-08" ...
```

56746

In [20]:

```
# remove na values and sort dates
defense_date <- defense_date [ ! is.na ( defense_date )]
defense_date <- sort ( defense_date )
```

In [21]:

```
# seperate year, month and day from date
defense_date_df <- data.frame ( defense_date )
defense_date_df <- defense_date_df %>% dplyr :: mutate ( year = lubridate :: y
head ( defense_date_df )
```

defense_date	year	month	day
1971-01-01	1971	1	1
1972-01-01	1972	1	1
1973-01-01	1973	1	1
1976-01-01	1976	1	1
1979-01-01	1979	1	1
1980-01-01	1980	1	1

In [22]:

```r
# select 1st of jan defense dates
jan_01_df  <-  defense_date_df  %>%  filter ( month  ==  1  &  day  ==  1 )
head ( jan_01_df ,  10 )
```

defense_date	year	month	day
1971-01-01	1971	1	1
1972-01-01	1972	1	1
1973-01-01	1973	1	1
1976-01-01	1976	1	1
1979-01-01	1979	1	1
1980-01-01	1980	1	1
1982-01-01	1982	1	1
1984-01-01	1984	1	1
1984-01-01	1984	1	1
1984-01-01	1984	1	1

In [23]:

```r
# get theses count for each year with jan 1st defence date
jan_01_df <- jan_01_df %>% select(year) %>% group_by(year) %>% count()
colnames(jan_01_df) <- c("Year", "Tot.Jan")
head(jan_01_df, 15)
```

Year	Tot.Jan
1971	1
1972	1
1973	1
1976	1
1979	1
1980	1
1982	1

Year	Tot.Jan
1984	6
1985	3007
1986	5162
1987	8439
1988	11045
1989	11102
1990	11011
1991	10831

In [24]:

```
# get total theses defended for each year
total_theses_df <- defense_date_df %>% select ( year ) %>% group_by ( year )
colnames ( total_theses_df ) <- c ( "Year" , "Tot.Year" )
head ( total_theses_df , 10 )
```

Year	Tot.Year
1971	1
1972	1
1973	1
1976	1
1979	1
1980	1
1982	1
1984	6
1985	3007
1986	5162

In [25]:

```
# get ratio column
jan_01_df <- inner_join ( jan_01_df , total_theses_df , by = 'Year' )
jan_01_df $ Portion.Theses <- jan_01_df $ Tot.Jan / jan_01_df $ Tot.Year
head ( jan_01_df )
```

Year	Tot.Jan	Tot.Year	Portion.Theses
1971	1	1	1
1972	1	1	1
1973	1	1	1
1976	1	1	1
1979	1	1	1

Year	Tot.Jan	Tot.Year	Portion.Theses
1980	1	1	1

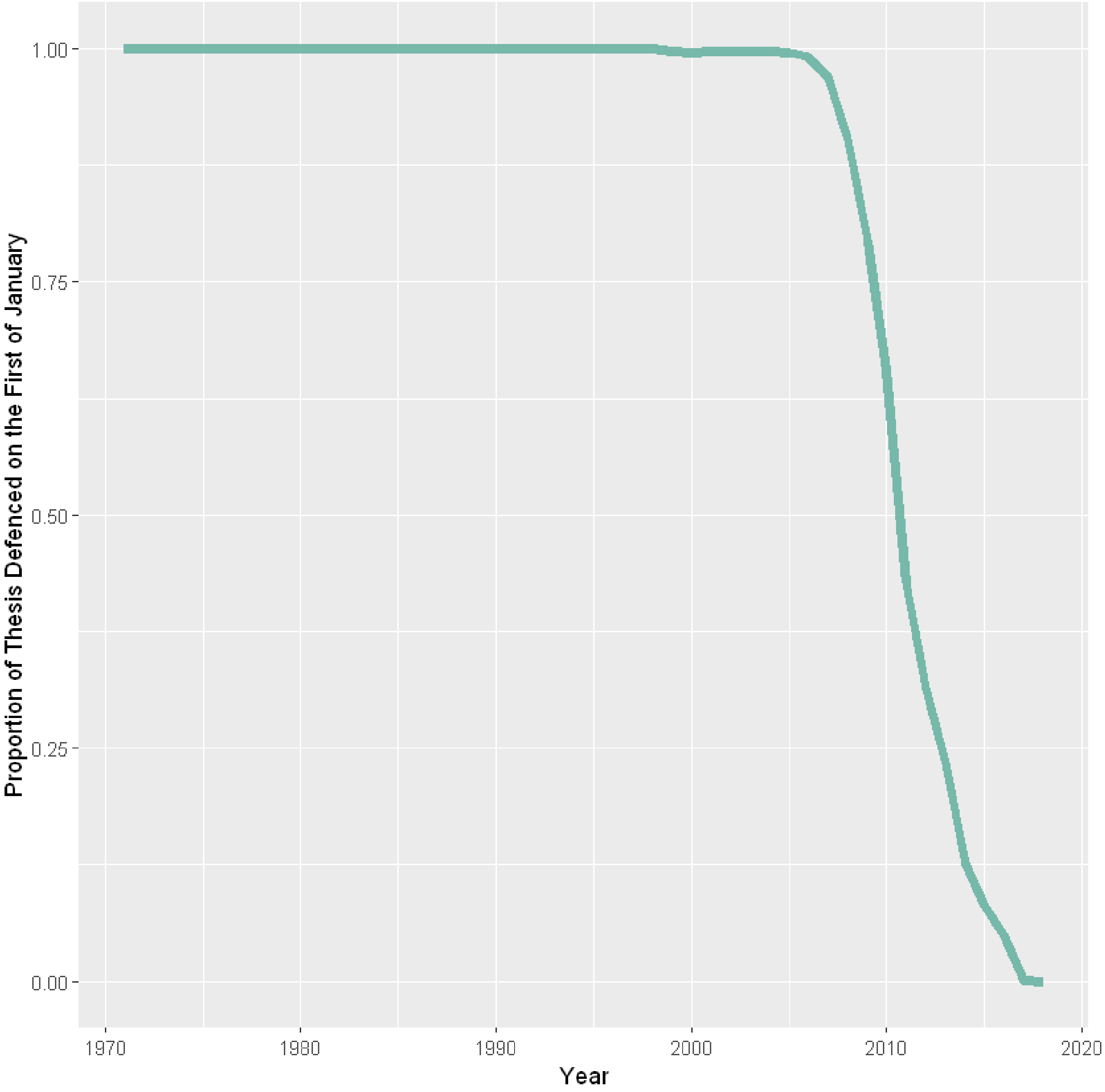
In [26]:

```
sum ( jan_01_df $ Tot.Jan )  /  sum ( jan_01_df $ Tot.Year )  *  100
```

74.1229597273658

In [27]:

```
# plot year vs ratio
ggplot ( jan_01_df ,  aes ( x = Year ,  y = Portion.Theses ))  +
    geom_line (  color = "# 69b3a2" ,  size = 2 ,  alpha = 0.9 )  +
    labs ( x  =  "Year" ,  y  =  "Proportion of Thesis Defenced on the First of Janu
```



In [28]:

```
# check for drop in ratio
subset ( jan_01_df ,  Year  <  2000 )
```

Year	Tot.Jan	Tot.Year	Portion.Theses
1971	1	1	1,000,000
1972	1	1	1,000,000

Year	Tot.Jan	Tot.Year	Portion.Theses
1973	1	1	1,000,000
1976	1	1	1,000,000
1979	1	1	1,000,000
1980	1	1	1,000,000
1982	1	1	1,000,000
1984	6	6	1,000,000
1985	3007	3007	1,000,000
1986	5162	5162	1,000,000
1987	8439	8439	1,000,000
1988	11045	11045	1,000,000
1989	11102	11102	1,000,000
1990	11011	11011	1,000,000
1991	10831	10831	1,000,000
1992	12065	12065	1,000,000
1993	12309	12309	1,000,000
1994	12991	12991	1,000,000
1995	10569	10569	1,000,000
1996	11354	11354	1,000,000
1997	11665	11669	0.9996572
1998	11015	11023	0.9992742
1999	10950	10982	0.9970861

In [29]:

```
subset ( jan_01_df ,  Year  >  2005  &  Year  <  2015 )
```

	Year	Tot.Jan	Tot.Year	Portion.Theses
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

	Year	Tot.Jan	Tot.Year	Portion.Theses
37	2013	3237	13868	0.2334151
38	2014	1666	13202	0.1261930

All thesis were defended on the 1st of Jan from 1971-1996 and slowly started to decrease. We can see a significant drop from 0.90 in 2008 to 0.79 in 2009.

> Check for author name homonyms + Cecile Martin

In [30]:

```
# select author and author id
Author_temp <- theses_df %>% select ( Author , Identifier.auteur ) %>% group_
Author_temp <- na.omit ( Author_temp )
colnames ( Author_temp) ) <- c ( "Author" , "ID" , "Freq" )
head ( Author_temp )
```

	Author	ID	Freq
2	Andrea Ramazzotti	174423705	2
80	Gilles Deshayes	182410528	1
135	Tuan Anh An Vo	190210486	1
616	Darine Chamsine	168134241	1
630	Liza Gladys Boukandou Kombila	189552883	1
819	Eve Duca	161896944	1

In [31]:

```
# get count for distinct author name and id pair
Author_temp <- Author_temp %>% arrange ( desc ( Freq ))
head ( Author_temp )
```

Author	ID	Freq
Catherine leport	69413916	7
Philippe Blanc	85924660	6
Thierry martin	60151013	6
Philippe Andre	61648493	5
Philippe Girard	61024228	5
Philippe Chevalier	66761999	5

In [32]:

```
# check for homonyms
all_author_temp <- Author_temp %>% select ( Author ) %>% group_by ( Author )
head ( all_author_temp )
homonym_temp <- all_author_temp %>% filter ( freq > 1 )
head ( homonym_temp )
```

Author	freq

Author	freq
#NAME?	1
(...) Massinga Kombila	1
. Aditya Arie Nugraha	1
. Edang Nnang	1
. Giang Tran Thi Hoang	1
. Govind	1

Author	freq
Abdallah Benaissa	2
Abdallah Dib	2
Abdallah Hiba	2
Abdelkader Mokhtari	2
Abdellatif El Hassani	2
Abdellatif Taghzouti	2

In [33]:

```
# compute portion of homonyms
nrow ( homonym_temp ) / nrow ( all_author_temp ) * 100
```

2.13960102648875

> If we consider all unique supervisor names in the dataset only 2% are homonymns

In [34]:

```
# analyze Cecile Martin case
subset ( Author_temp , Author_temp $ Author == "Cecile Martin" )
```

	Author	ID	Freq
39	Cecile Martin	81323557	4
36564	Cecile Martin	203208145	1
161937	Cecile Martin	179423568	1
273584	Cecile Martin	182118703	1

> Issues in the supervisor's ID

In [ ]:

```
unique ( theses_df $ Identifier.director )
```

In [36]:

```
# get length of supervisor id for each thesis
director_id <- theses_df $ Identifier.director
id_temp_01 <- data.frame ( director_id )
id_temp_01 <- na.omit ( id_temp_01 )
```

```
id_temp_01 $ director_id  <-  as.character ( id_temp_01 $ director_id )
id_temp_01 $ length  <-  nchar ( id_temp_01 $ director_id )
head ( id_temp_01 )
```

	director_id	length
1	29561248	8
2	715,441,511	11
3	57030758	8
6	26941848	8
8	34508287	8
9	32574088	8

In [37]:

```
# get fequency of each length
id_len_temp_01  <-  id_temp_01  %>%  select ( length )  %>%  group_by ( length )  %>
xtable ( id_len_temp_01 )
```

length	freq
1	4587
2	137
8	255680
9	78960
11	59108

In [38]:

```
# get percentage of each length
total_id  <-  sum ( id_len_temp_01 $ freq )
id_len_temp_01 $ portion  <-  ( id_len_temp_01 $ freq  /  total_id )  *  100
xtable ( id_len_temp_01 )
```

length	freq	portion
1	4587	1.15114738
2	137	0.03438134
8	255680	64.16511072
9	78960	19.81569596
11	59108	14.83366460

In [39]:

```
# ids with comma
director_temp_01  <-  filter ( theses_df ,  grepl ( "," ,  theses_df $ Identifier.di
nrow ( director_temp_01 )
head ( director_temp_01 )
```

59108

Author	Author ID	Title	Supervisor	Director.of.thesis..name.firstname.	Manager ID	S
Andrea Ramazzotti	174423705	Application of the PGD to the resolution of transient couples problems with a view to the lightening of composite structures.	Jean-Claude Grandidier, Marianne Beringhier	Grandidier Jean-Claude, Beringhier Marianne	715,441,511	
Ioana Raluca Andreescu	N / A	Robinson in the Ile de la Pape. Representations of the social system in post-war Ilian European literature	Annick Louis, Jean-Louis Fabiani	Louis Annick, Fabiani Jean-Louis	348,740,620	
Tarik Khoutaif	N / A	Study and modeling of synchronous bluetooth links for an architecture of real-time communicating systems.	Thierry Val, Fabrice Peyrard	Val Thierry, Peyrard Fabrice	113,464,657	
Guilhem Armand	N / A	Fictions with a scientific vocation from Cyrano de Bergerac to Diderot: towards a hybrid poetry	Jean-Michel Racault, Aurelia Gaillard	Racault Jean-Michel, Gaillard Aurelia	283,003,190	
Aman ghelich Atabaei	N / A	Interbank market and contagions in times of financial crisis.	Daniel Goyeau, Catherine Lubochinsky	Goyeau Daniel, Lubochinsky Catherine	562,440,960	

Author	Author ID	Title	Supervisor	Director.of.thesis..name.firstname.	Manager ID	S
Samuel Brosset	N / A	The contexts of interaction and integration of Icelandic information networks, between rhythm, constraint and identity, its limits and its exemplarity.	Catherine Bernie-Boissard, Dominique Crozat	Bernie-Boissard Catherine, Crozat Dominique	327,131,260	

In [40]:

```
# thesis with 2 supervisors
nrow ( filter ( director_temp_01 ,  grepl ( "," ,  Director.thesis ,  fixed  =  TRUE
```

59108

In [41]:

```
# percentage of thesis with more than 1 supervisor
nrow ( filter ( director_temp_01 ,  grepl ( "," ,  Director.these ,  fixed  =  TRUE
```

62.1182504151165

62.12% of the time if the theses has more than supervisor the supervisor id has a comma.

In [98]:

```
# ids with X
director_temp_02  <-  filter ( theses_df ,  grepl ( "X" ,  Identifier.director ,  fi
nrow ( director_temp_02 )
```

35163

In [43]:

```
# percentage of thesis with more than 1 supervisor
nrow ( filter ( director_temp_02 ,  grepl ( "," ,  Director.these ,  fixed  =  TRUE
```

7.85989028312

> 7.86% of the time if the theses has more than 1 supervisor the supervisor id has a X.

In [44]:

```
# percentage of thesis with 1 supervisor
( nrow ( director_temp_02 )  -  nrow ( filter ( director_temp_02 ,  grepl ( "," ,  D
```

7.85383982524327

> 7.85% of the time if the theses has 1 supervisor the supervisor id has a X.

> Number of PHD defended over the years

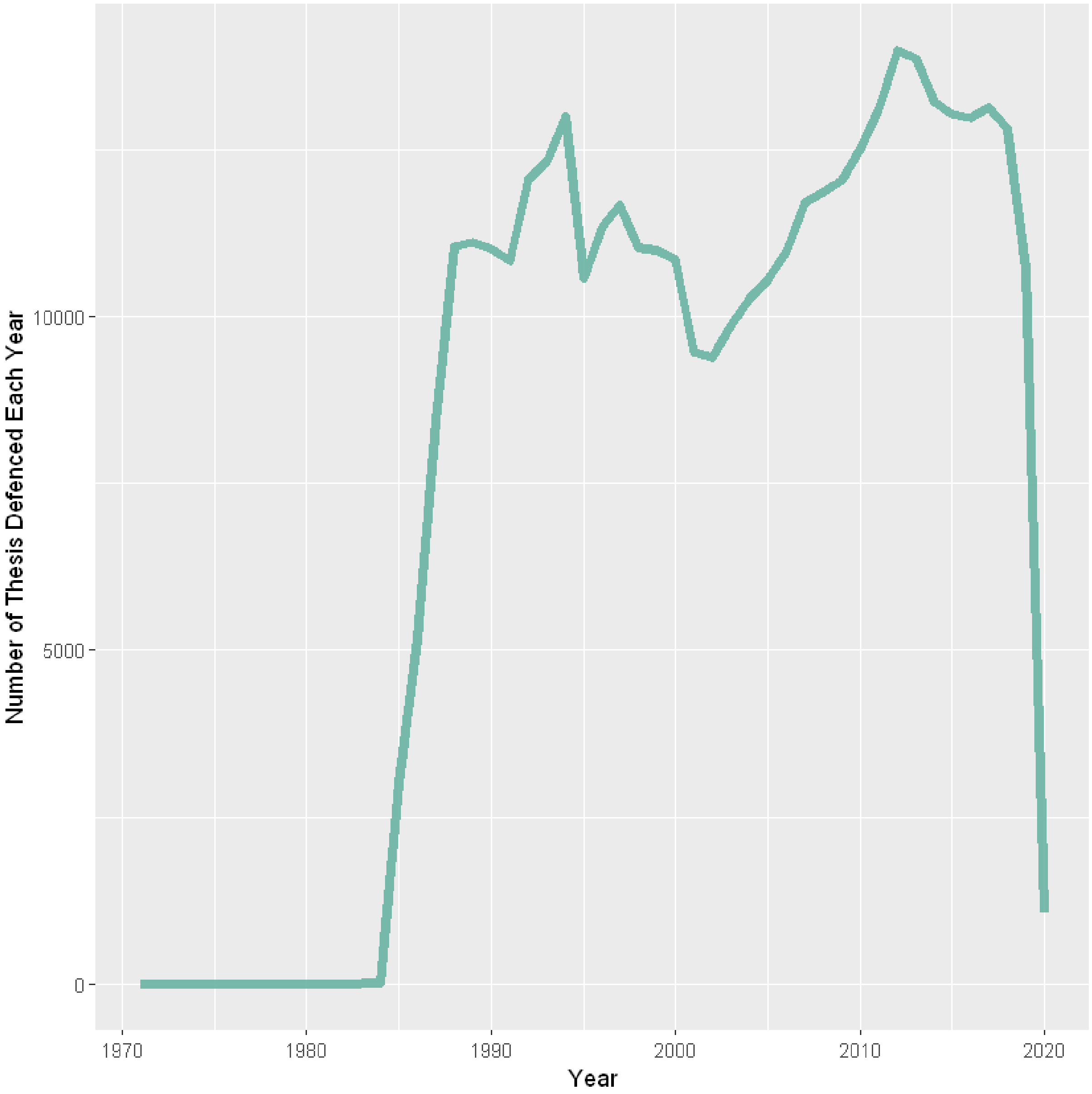
In [45]:

```
head ( total_theses_df ,  10 )
```

Year	Tot.Year
1971	1
1972	1
1973	1
1976	1
1979	1
1980	1
1982	1
1984	6
1985	3007
1986	5162

In [46]:

```
# plot trend of num of theses defended
ggplot ( total_theses_df ,  aes ( x = Year ,  y = Tot.Year ))  +
    geom_line ( color = "# 69b3a2" ,  size = 2 ,  alpha = 0.9 )  +
    labs ( x  =  "Year " ,  y  =  " Number of Thesis Defenced Each Year " )
```



In [47]:

```
# verify rise year
subset ( total_theses_df ,  Year  >  1980  &  Year  <=  1990 )
```

	Year	Tot.Year
7	1982	1
8	1984	6
9	1985	3007
10	1986	5162
11	1987	8439
12	1988	11045
13	1989	11102
14	1990	11011

In [48]:

```
# verify drop is in 2019-2020
subset ( total_theses_df ,  Year  >  2015  &  Year  <=  2020 )
```

	Year	Tot.Year
40	2016	12965
41	2017	13123
42	2018	12805
43	2019	10712
44	2020	1070

We see that there is a sudden drop in the number of PHDs defended from 10712 in 2019 to 1070 in 2020. This might be for the following reasons:

- The Covid-19 lockdown

-

3.4. Outliers

Supervisor

In [49]:

```
# Unnest rows with multiple supervisors
df_unnest <- theses_df %>% unnest ( Director = strsplit ( tolower ( Director.t
head ( df_unnest )
```

```
Warning message:
"unnest () has a new interface. See? unnest for details.
Try `df%>% unnest (c (Director))`, with `mutate ()` if needed "
```

Author	Author ID	Title	Supervisor	Director.of.thesis..name.firstname.	Manag
Saeed al marri	N / A	Documentary credit and the enforceability of exceptions	Philippe Delebecque	Delebecque Philippe	295612
Andrea Ramazzotti	174423705	Application of the PGD to the resolution of transient couples problems with a view to the lightening of composite structures.	Jean-Claude Grandidier, Marianne Beringhier	Grandidier Jean-Claude, Beringhier Marianne	715,441,5

Author	Author ID	Title	Supervisor	Director.of.thesis..name.firstname.	Manag
Andrea Ramazzotti	174423705	Application of the PGD to the resolution of transient couples problems with a view to the lightening of composite structures.	Jean-Claude Grandidier, Marianne Beringhier	Grandidier Jean-Claude, Beringhier Marianne	715,441,5
OLIVIER BODENREIDER	N / A	Design of a computer tool for the study of kinetics observed in clinical toxicology	Francois Kohler	Kohler Francois	570307
Emmanuel Porte	N / A	Socio-history of public policies in social matters concerning students.	Gilles Pollet	Pollet Gilles	N /
Arthur devriendt	N / A	INFORMATION AND COMMUNICATION TECHNOLOGIES AND NEW RURALITIES.	Gabriel Dupuy	Dupuy Gabriel	N /

In [50]:

```
# get freq of unique director name & id pair
df_directeur <- df_unnest %>% select ( Director , Identifier.director ) %>% gr
head ( df_director , 20 )
```

Director	Manager ID	freq
philippe delebecque	29561248	178
jean-claude grandidier	715,441,511	5
marianne beringhier	715,441,511	3
francois kohler	57030758	12
gilles pollet	N / A	7
gabriel dupuy	N / A	2
edmond jouve	26941848	46
stone count	N / A	3

Director	Manager ID	freq
laurent sermet	34508287	5
anne-emmanuelle berger	32574088	4
jean-pierre keyboard	35557060	13
patrice vermeren	28251873	32
jerome julien	N / A	4
deen gibirila	33883238	20
danielle cabanis	N / A	6
jean-michel ganteau	58596852	11
emile-henri riard	137391919	9
serge regourd	27093115	46
bernard boene	27093115	1
elisabeth claverie	76120333	5

In [51]:

```
# drop na values in director name and sort
df_directeur [ df_directeur  ==  "" ]  <-  NA
df_directeur [ df_directeur  ==  "" ]  <-  NA
df_directeur  <-  df_directeur  %>%  drop_na ()  %>%  arrange ( desc ( freq ) )
xtable ( head ( df_director ,  20 ))
```

Director	Manager ID	freq
jean-michel scherrmann	59375140	208
francois-paul blanc	26730774	205
pierre brunel	26756625	193
philippe delebecque	29561248	178
guy pujolle	27084868	177
michel bertucat	98531891	173
bernard teyssie	27158578	146
bruno foucart	26870177	132
henry de lumley	26997894	132
jean-claude chaumeil	58552499	131
michel maffesoli	27001067	128
roger g. boulu	59209143	127
daniel-henri pageaux	02705554X	124
georges molinie	02703352X	116

Director	Manager ID	freq
jean bessiere	26725916	114
francis balle	26702606	109
gregoire loiseau	35137576	101
michel meslin	27024938	96
eliane chiron	26787083	96
pierre-philippe rey	55477046	96

> Quantiles to find Outliers

In [53]:

```
# get lower bound
lower_bound <- quantile(df_directeur$freq, 0.01)
lower_bound
```

1%: 1

In [54]:

```
# get upper bound
upper_bound <- quantile(df_directeur$freq, 0.997)
upper_bound
```

99.7%: 35

> more than 35 is outlier

Author

In [56]:

```
head(Auteur_temp)
```

Auteur	ID	Freq
Catherine Leport	69413916	7
Philippe Blanc	85924660	6
Thierry Martin	60151013	6
Philippe Andre	61648493	5
Philippe Girard	61024228	5
Philippe Chevalier	66761999	5

In [57]:

```
# get lower bound
lower_bound <- quantile(Auteur_temp$Freq, 0.01)
lower_bound
```

1%: 1

In [58]:

```
# get upper bound
upper_bound  <-  quantile ( Author_temp $ Freq ,  0.997 )
upper_bound
```

99.7%: 2

> more than 2 is outlier

3.5. Perliminary Results

1) **Language**

In [60]:

```
# get theses language
languages_df  <-  theses_df $ Langue.de.la.these
head ( languages_df )
```

1. N / A
2. N / A
3. 'Fr'
4. N / A
5. N / A
6. N / A

In [61]:

```
# sort and set to lower case
languages_df  <-  na.omit ( languages_df )
languages_df  <-  sort ( languages_df )
languages_df  <-  data.frame ( languages_df )
colnames ( languages_df )  <-  c ( "language" )
languages_df $ language  <-  tolower ( languages_df $ language )
head ( languages_df )
```

language
aafr
aafr
aafr
ab
ab
abfr

In [62]:

```
# get number of languages
languages_df $ n.language  <-  str_length ( languages_df $ language )  /  2
head ( languages_df )
```

language	n.language
aafr	2
aafr	2
aafr	2
ab	1

language	n.language
ab	1
abfr	2

In [63]:

```
# categorize language
languages_df <- languages_df %>% mutate ( lang.type = case_when (
    ( n.language == 1 & language == "en" ) ~ "English" ,
    ( n.language == 1 & language == " fr " ) ~ " French " ,
    ( n.language == 2 & ( str_contains ( language , " en ") == TRUE | ) == F
                          n.language > 2 )) ~ "Other" ,
    ))
head ( languages_df )
tail ( languages_df )
```

language	n.language	lang.type
aafr	2	Bilingual
aafr	2	Bilingual
aafr	2	Bilingual
ab	1	Other
ab	1	Other
abfr	2	Bilingual

	language	n.language	lang.type
383874	zhfr	2	Bilingual
383875	zhfr	2	Bilingual
383876	zhfr	2	Bilingual
383877	zhfr	2	Bilingual
383878	zhfrit	3	Other
383879	zhfrug	3	Other

In [64]:

```
# get count
lang_type_df <- languages_df %>% select ( lang.type ) %>% group_by ( lang.type
lang_type_df
```

lang.type	freq
Bilingual	16488
English	30942
French	334406
Other	2043

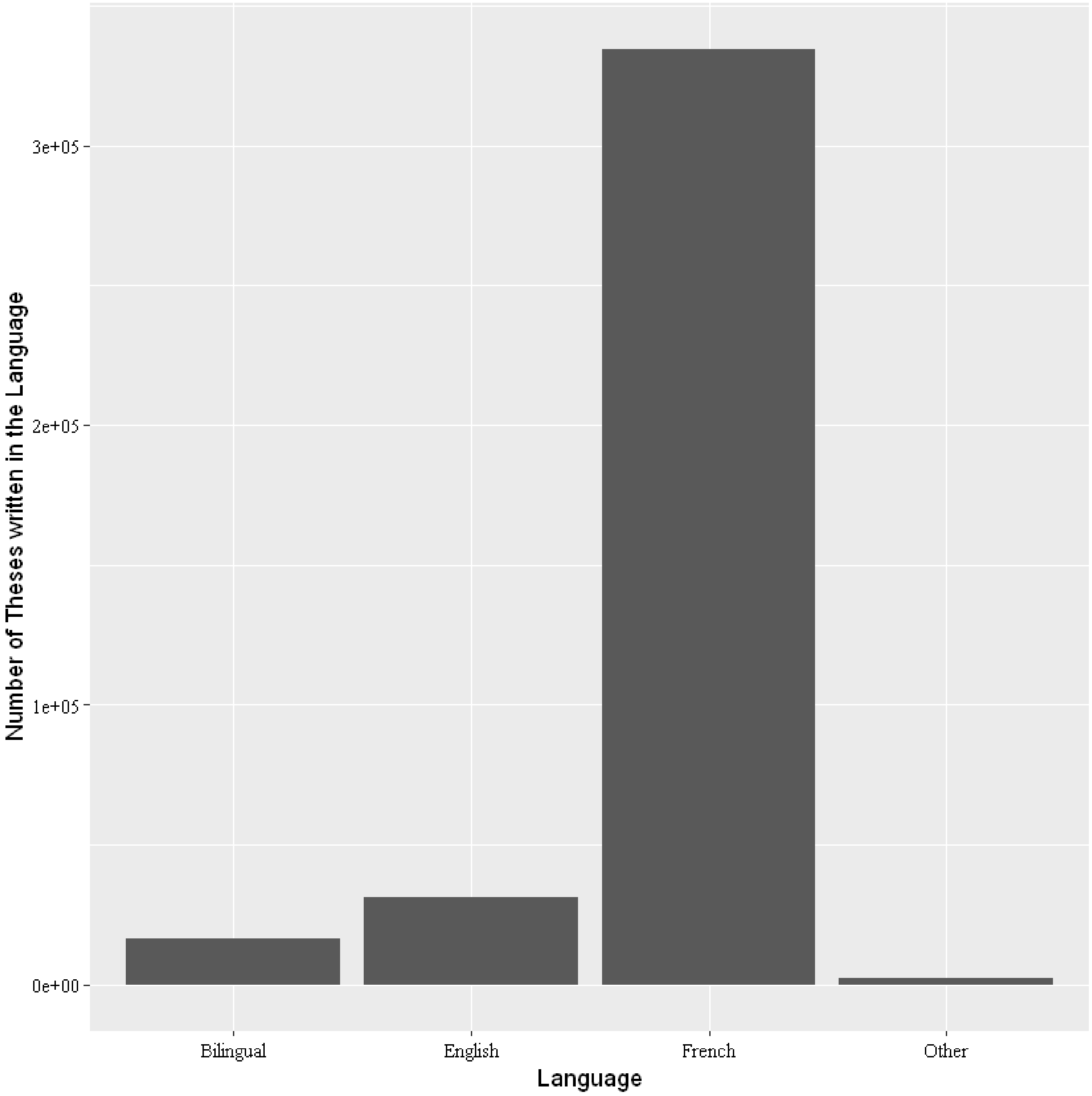
In [65]:

```
# set levels
lang_type_df $ lang.type <-  factor ( lang_type_df $ lang.type ,  levels  =  lang_t
lang_type_df
```

lang.type	freq
Bilingual	16488
English	30942
French	334406
Other	2043

In [66]:

```
# plot total number for each lang
ggplot ( lang_type_df ,  aes ( x = lang.type ,  y = freq ))  +
geom_bar ( stat  =  "identity" )  +
theme ( plot.title  =  element_text ( family  =  "serif" ,  color  =  "black" ),
    axis.text.x  =  element_text ( family  =  "serif" ,  color  =  "black" ),
    axis.text.y  =  element_text ( family  =  "serif" ,  color  =  "black" ))  +
labs ( x  =  "Language" ,  y  =  "Number of Theses written in the Language" )
```



In [67]:

```
# select language and defense date
```

```
df_lang_date  <-  theses_df  %>%  select ( Date.de.soutenance , Langue.de.these )  %
colnames ( df_lang_date )  <-  c ( "defense.date" ,  " language " )
head ( df_lang_date )
```

defense.date	language
1993-01-01	Fr
2015-01-01	Fr
2015-01-01	Fr
2013-12-07	Fr
2013-11-25	Fr
2013-11-22	Fr

In [68]:

```
# get num of languages and set to lower case
df_lang_date $ n.language  <-  str_length ( df_lang_date $ language )  /  2
df_lang_date $ language  <-  tolower ( df_lang_date $ language )

# get year of defense
df_lang_date  <-  df_lang_date  %>%  dplyr :: mutate ( year  =  lubridate :: year (
df_lang_date  <-  df_lang_date [ order ( df_lang_date $ year ),]
head ( df_lang_date )
```

defense.date	language	n.language	year
1971-01-01	Fr	1	1971
1972-01-01	Fr	1	1972
1973-01-01	Fr	1	1973
1976-01-01	Fr	1	1976
1979-01-01	Fr	1	1979
1980-01-01	Fr	1	1980

In [69]:

```
# Add if theses is:
# French
# English
# Bilingual -> theses done in french or english and one other language
# Other

df_lang_date  <-  df_lang_date  %>%  mutate ( lang.type  =  case_when (
    ( n.language  ==  1  &  language  ==  "en" )  ~  "English" ,
    ( n.language  ==  1  &  language  ==  "fr" )  ~  "French" ,
    ( n.language  ==  2  &  ( str_contains ( language ,  "en" )  ==  TRUE  |  str_co
    (( n.language  ==  1  &  language  ! =  "en"  &  language  ! =  "fr" )  |  ( n.l
    ))
head ( df_lang_date )
tail ( df_lang_date )
```

defense.date	language	n.language	year	lang.type
1971-01-01	Fr	1	1971	French

defense.date	language	n.language	year	lang.type
1972-01-01	Fr	1	1972	French
1973-01-01	Fr	1	1973	French
1976-01-01	Fr	1	1976	French
1979-01-01	Fr	1	1979	French
1980-01-01	Fr	1	1980	French

defense.date	language	n.language	year	lang.type
2020-01-10	Fr	1	2020	French
2020-06-26	fren	2	2020	Bilingual
2020-02-06	in	1	2020	English
2020-06-11	Fr	1	2020	French
2020-05-07	Fr	1	2020	French
2020-06-23	Fr	1	2020	French

> Filter out data before 1985 because of sudden rise and 2020 because we don't have data for the whole year

In [70]:

```
# select between 1988 and 2020
df_lang_date <- df_lang_date %>% filter ( year > 1988 & year < 2020 )
```

In [71]:

```
# get count of theses done in each language type each year
df_lang_type_ts <- df_lang_date %>% select ( year , lang.type ) %>% group_by
colnames ( df_lang_type_ts ) <- c ( "Year" , "Lang_Type" , "Sum.Lang_Type" )
head ( df_lang_type_ts )
tail ( df_lang_type_ts )
```

Year	Lang_Type	Sum.Lang_Type
1989	Bilingual	234
1989	English	4
1989	French	10860
1989	Other	3
1990	Bilingual	206
1990	English	12

	Year	Lang_Type	Sum.Lang_Type
118	2018	French	7807
119	2018	Other	122

	Year	Lang_Type	Sum.Lang_Type
120	2019	Bilingual	550
121	2019	English	2818
122	2019	French	5615
123	2019	Other	95

In [72]:

```r
# get count of theses done each year
df_year <- df_lang_date %>% select ( year ) %>% group_by ( year ) %>% count
colnames ( df_year ) <- c ( "Year" , "Sum.Year" )
head ( df_year )
```

Year	Sum.Year
1989	11101
1990	11011
1991	10831
1992	12064
1993	12308
1994	12991

In [73]:

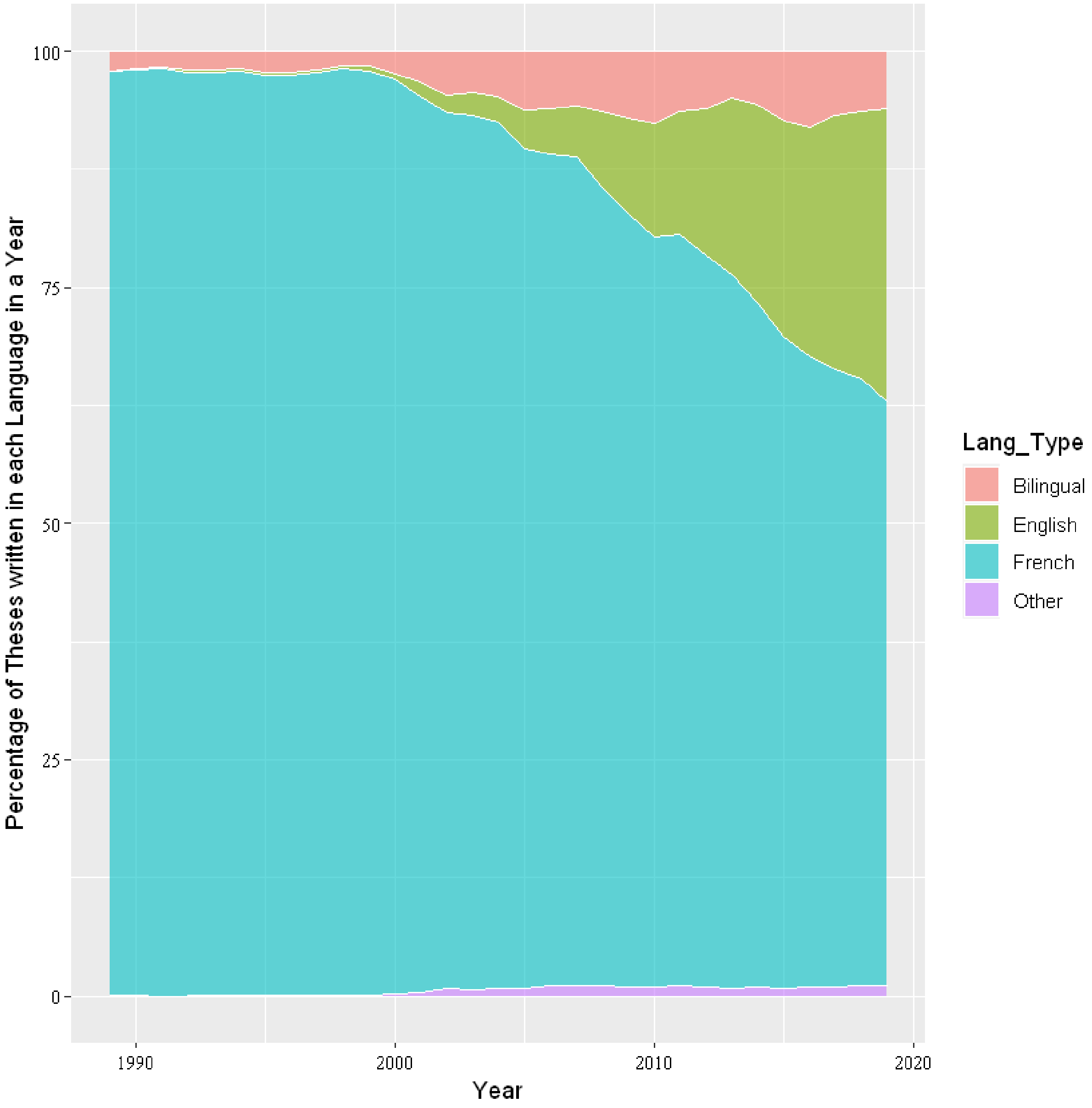
```r
# Get the percentage of theses done in Each language kind Each year
full_lang_type <- full_join ( df_lang_type_ts , df_year , by = 'Year' )
full_lang_type $ Sum.Percentage <- round (( full_lang_type $ Sum.Lang_Type / ful
head ( full_lang_type )
```

Year	Lang_Type	Sum.Lang_Type	Sum.Year	Sum.Percentage
1989	Bilingual	234	11101	2.11
1989	English	4	11101	0.04
1989	French	10860	11101	97.83
1989	Other	3	11101	0.03
1990	Bilingual	206	11011	1.87
1990	English	12	11011	0.11

> ggplot2 graph

In [74]:

```r
# ggplot2 for percentage
ggplot ( full_lang_type , aes ( x = Year , y = Sum.Percentage , fill = Lang_Type
geom_area ( alpha = 0.6 , size = . 5 , color = "white" ) +
theme ( plot.title = element_text ( family = "serif" , color = "black" ),
    axis.text.x = element_text ( family = "serif" , color = "black" ),
    axis.text.y = element_text ( family = "serif" , color = "black" )) +
labs ( y = "Percentage of Theses written in each Language in a Year " )
```



> plotly graph

In [75]:

```
# plotly for percentage
fig <- plot_ly (
    type = 'scatter' ,
    x = full_lang_type $ Year ,
    y = full_lang_type $ Sum.Percentage ,
    color = full_lang_type $ Lang_Type ,
    mode = "lines" ,
    fill = "tozeroy"
)
fig <- fig %>% layout ( xaxis = list ( title = 'Year' , layout.font = "Tim
         yaxis = list ( title = 'Percentage of Theses done in each Language in a

fig
```

In [76]:

```
# plotly for percentage
p  <-  full_lang_type  %>%  ggplot ( aes ( x = Year ,  y = Sum.Percentage ,  fill =
    geom_area ( alpha = 0.6  ,  size = . 4 ,  color = "gray" )   +
theme ( plot.title  =  element_text(family = "serif", color = "black"),
    axis.text.x = element_text(family = "serif", color = "black"),
    axis.text.y = element_text(family = "serif", color = "black")) +
    labs(y = "Percentage of Theses done in each Language")
ggplotly(p, tooltip="text")
```

2) **Period of Year**

In [78]:

```
# to select defense between 1996 to 2020
defense_date_df  <-  defense_date_df  %>%  filter ( year  >  1996  &  year  <  2020
```

In [79]:

```
# select january and remove jan 1
no_jan_01_df  <-  defense_date_df  %>%  filter ( month  ==  1  &  day  ! =  1 )
head ( no_jan_01_df ,  10 )
```

defense_date	year	month	day
1999-01-09	1999	1	9
1999-01-13	1999	1	13
1999-01-14	1999	1	14
1999-01-19	1999	1	19
1999-01-19	1999	1	19
1999-01-21	1999	1	21
2000-01-12	2000	1	12
2000-01-17	2000	1	17
2000-01-21	2000	1	21
2000-01-24	2000	1	24

In [81]:

```
# select months except jan
no_jan_df  <-  defense_date_df  %>%  filter ( month  ! =  1 )
head ( no_jan_df )
```

defense_date	year	month	day
1997-03-29	1997	3	29
1997-09-19	1997	9	19
1997-12-01	1997	12	1
1997-12-06	1997	12	6
1998-12-03	1998	12	3
1998-12-09	1998	12	9

In [82]:

```
# merge
no_jan_01_df  <-  rbind.fill ( no_jan_01_df ,  no_jan_df )
```

In [83]:

```
# get month count for each year
dates_df_ym  <-  no_jan_01_df  %>%  select ( year ,  month )  %>%  group_by ( year ,
colnames ( dates_df_ym )  <-  c ( "Year" ,  "Month" ,  "Sum.Month" )
```

In [84]:

```
# get total count for each year
dates_df_y  <-  no_jan_01_df  %>%  select ( year )  %>%  group_by ( year )  %>%  cou
colnames ( dates_df_y )  <-  c ( "Year" ,  "Sum.Year" )
```

In [85]:

```
# merge and get percentage
full_date  <-  full_join ( dates_df_ym ,  dates_df_y ,  by  =  'Year' )
full_date $ Sum.Percentage  <-  round (( full_date $ Sum.Month  /  full_date $ Sum.Y
head ( full_date )
```

Year	Month	Sum.Month	Sum.Year	Sum.Percentage
1997	3	1	4	25.00
1997	9	1	4	25.00
1997	12	2	4	50.00
1998	12	8	8	100.00
1999	1	6	32	18.75
1999	4	1	32	3.12

In [86]:

```
# get mean and sd
date_summary <- ddply ( full_date , ~ Month , summarize , mean = mean ( Sum.P

# change month from num to text and create levels
date_summary <- date_summary %>% mutate ( Month = month.name [ Month ])
date_summary $ Month <- factor ( date_summary $ Month , levels = date_summary $
```

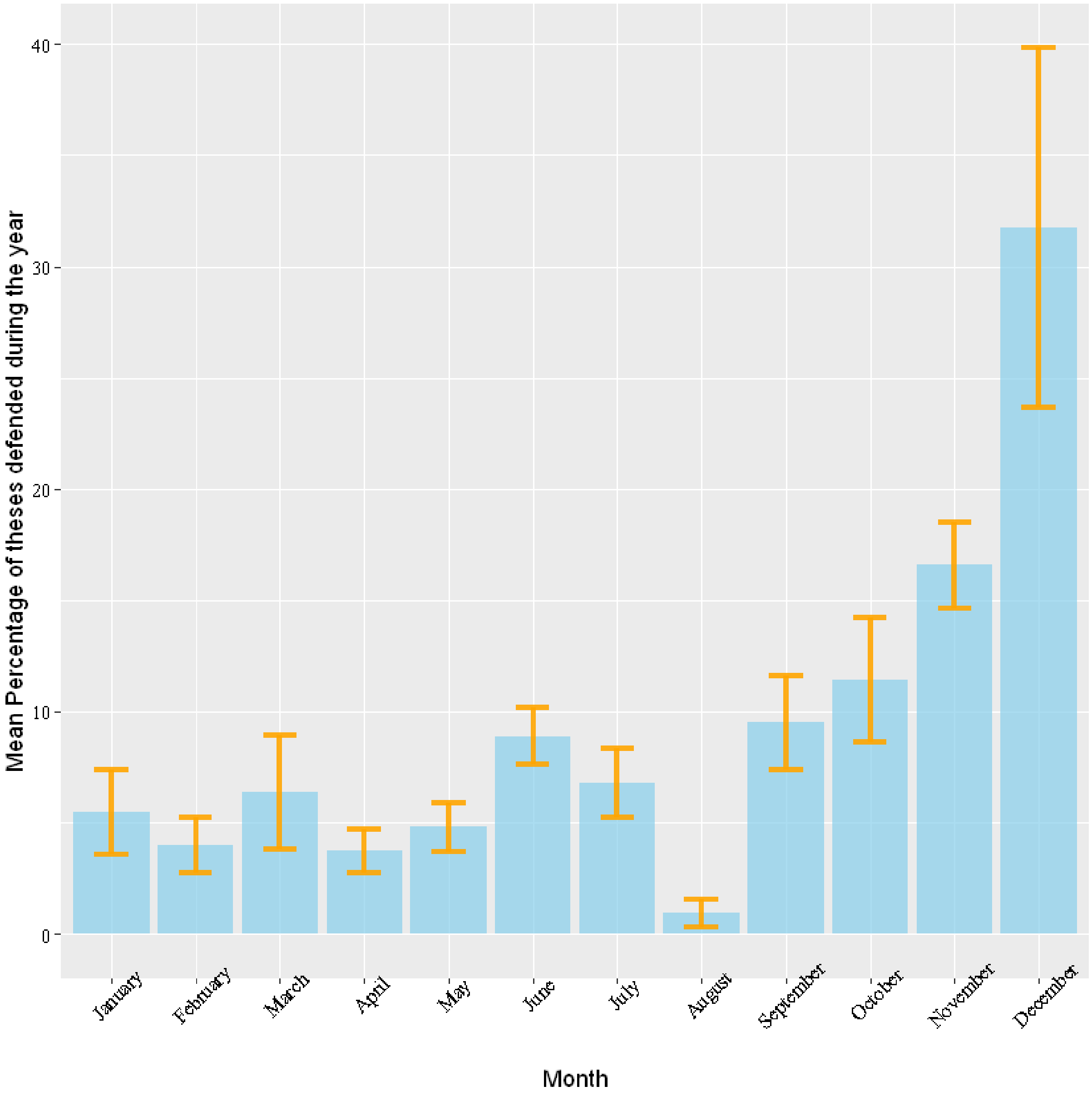
In [87]:

```
date_summary
```

Month	mean	sd	sd_02
January	5.4790000	3.777090	1.8885450
February	3.9976471	2.529418	1.2647091
March	6.3773684	5.133847	2.5669236
April	3.7266667	1.966930	0.9834648
May	4.8000000	2.173694	1.0868469
June	8.8915000	2.541956	1.2709779
July	6.7942857	3.121983	1.5609913
August	0.9413333	1.271382	0.6356910
September	9.5025000	4.198066	2.0990329
October	11.4119048	5.574342	2.7871709
November	16.5880952	3.856125	1.9280623
December	31.7504348	16.166461	8.0832307

In [88]:

```
# Plot
ggplot ( date_summary ) +
geom_bar ( aes ( x = Month , y = mean ), stat = "identity" , fill = "skyblue" ,
geom_errorbar ( aes ( x = Month , ymin = mean - sd_02 , ymax = mean + sd_02 ), wi
theme ( plot.title = element_text ( family = "serif" , color = "black" ),
      axis.text.x = element_text ( family = "serif " , color = " black " , an
     axis.text.y = element_text ( family = "serif" , color = "black" )) +
labs ( x = "Month" , y = "Mean Percentage of theses defended during the year" )
```



3) **Gender**

> Done in Python

4) **Bonus**

In [89]:

```
# select establishment name id and defense date
establishment_df  <-  theses_df [ c ( 'Etablissement.de.soutenance' ,  'Identifier.e
establishment_df  <-  na.omit ( establishment_df )

# get year of defense
establishment_df  <-  establishment_df  %>%  dplyr :: mutate ( Year  =  lubridate ::
head ( establishment_df )
```

Support.Establishment	Institution.identifier	Support date	Year
Paris 5	26404788	2008-11-24	2008
Saint Etienne	28209966	2005-07-01	2005
The meeting	26404451	2009-12-08	2009
Paris 8	26403552	2013-01-10	2013

Support.Establishment	Institution.identifier	Support date	Year
Nantes	26403447	2011-06-24	2011
Paris 8	26403552	2010-11-26	2010

In [90]:

```
# get yearly theses count for each uni
establishment_cpt_df  <-  establishment_df  %>%  select ( Year ,  Etablissement.de.s
head ( establishment_cpt_df )
```

Year	Support.Establishment	Institution.identifier	freq
1973	University of Nancy I	26403390	1
1979	Paris 10	26403587	1
1980	Nice	26403498	1
1982	Paris 4	26403633	1
1984	Limoges	26403315	1
1984	Mulhouse	26403250	3

In [91]:

```
# get yearly total
date_cpt_df  <-  establishment_df   %>%  select ( Year )  %>%  group_by ( Year )  %>
head ( date_cpt_df )
```

Year	freq
1973	1
1979	1
1980	1
1982	1
1984	6
1985	2987

In [92]:

```
# merge and get percentage
full_establishment_df  <-  full_join ( establishment_cpt_df ,  date_cpt_df ,  by  =
full_establishment_df $ Percentage  <-  full_establishment_df $ freq.x  /  full_esta
head ( full_establishment_df )
```

Year	Support.Establishment	Institution.identifier	freq.x	freq.y	Percentage
1973	University of Nancy I	26403390	1	1	100,000,000
1979	Paris 10	26403587	1	1	100,000,000
1980	Nice	26403498	1	1	100,000,000
1982	Paris 4	26403633	1	1	100,000,000
1984	Limoges	26403315	1	6	16.66667

Year	Support.Establishment	Institution.identifier	freq.x	freq.y	Percentage
1984	Mulhouse	26403250	3	6	50.00000

In [93]:

```
# get total theses count for each uni
establishment_cpt_df_2  <-  establishment_df  %>%  select ( Etablissement.de.soutena
head ( establishment_cpt_df_2 )
```

Support.Establishment	Institution.identifier	freq
[Amiens], University of Picardy - Jules Verne, Doctoral school in human and social sciences	26403714	1
[Amiens], University Picardie - Jules Verne, Doctoral School of Letters and Human Sciences, Department of Economics and Management	26403714	1
[Grenoble INPG]	26388804	1
AgroParisTech	139408088	65
Aix en Provence	26403781	1
Aix-Marseille	67331149	1

In [94]:

```
# sort
establishment_cpt_df_2  <-  establishment_cpt_df_2  %>%  arrange ( desc ( freq ))
head ( establishment_cpt_df_2 )
```

Support.Establishment	Institution.identifier	freq
Paris 6	27787087	20914
Paris 11	26404664	15326
Paris 7	27542084	11075
Paris 1	27361802	10749
Toulouse 3	26404672	9554
Paris 4	26403633	8277

In [95]:

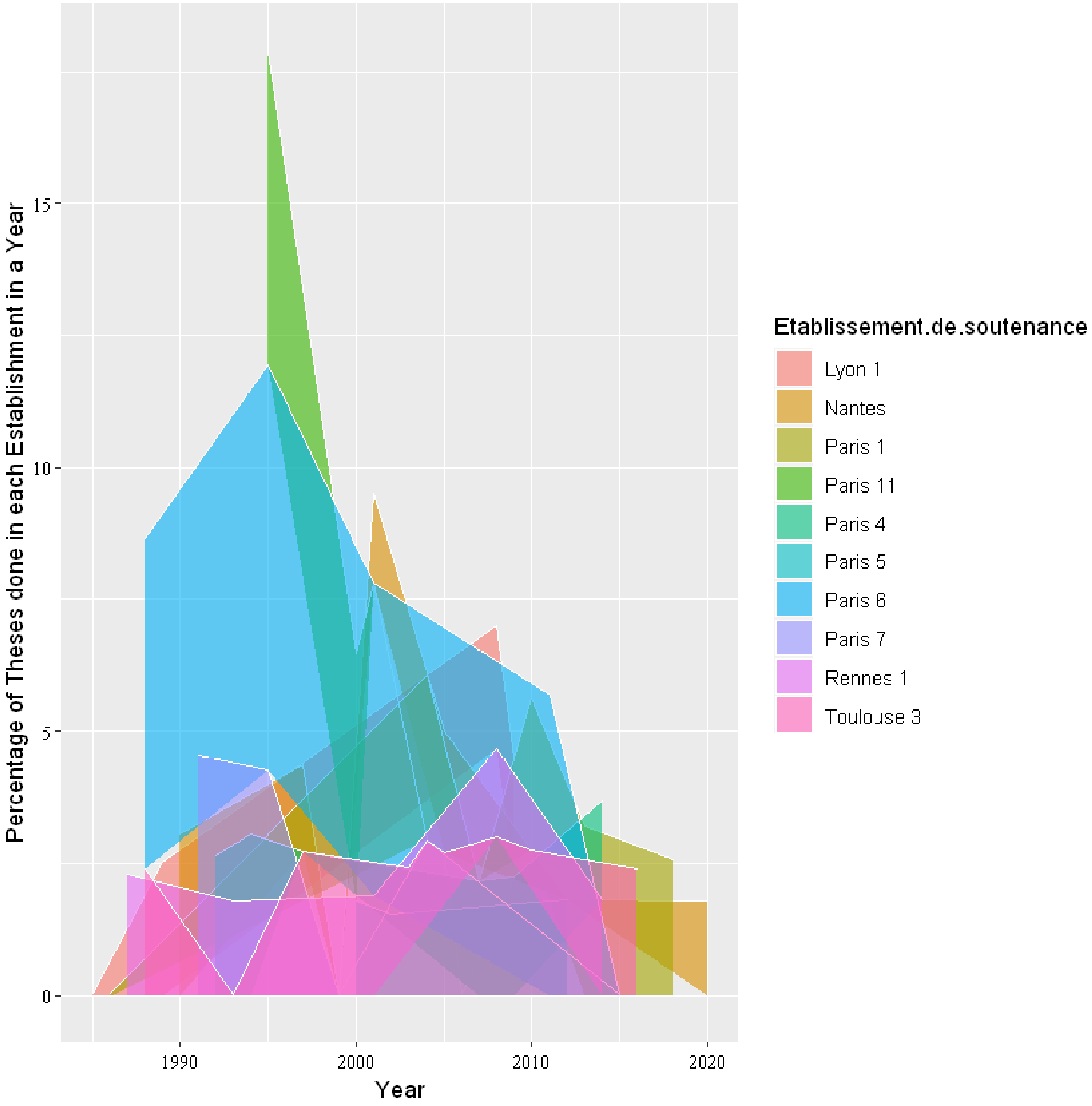
```
# get to 10 with highest theses count
highest_establishment  <-  establishment_cpt_df_2 [ 1 : 10 ,  "Etablissement.de.sout
highest_establishment
highest_id  <-  which ( full_establishment_df $ Etablissement.de.soutenance  ==   hi
highest_establishment_df  <-  full_establishment_df [ highest_id ,  ]
head ( highest_establishment_df )
```

1. 'Paris 6'
2. 'Paris 11'
3. 'Paris 7'
4. 'Paris 1'
5. 'Toulouse 3'
6. 'Paris 4'

7. 'Lyon 1'
8. 'Paris 5'
9. 'Nantes'
10. 'Rennes 1'

In [97]:

```
# plot
ggplot ( highest_establishment_df ,  aes ( x = Year ,  y = Percentage ,  fill = Etab
geom_area ( alpha = 0.6  ,  size = . 5 ,  color = "white" )  +
theme ( plot.title  =  element_text ( family  =  "serif" ,  color  =  "black" ),
    axis.text.x  =  element_text ( family  =  "serif" ,  color  =  "black" ),
    axis.text.y  =  element_text ( family  =  "serif" ,  color  =  "black" ))  +
labs ( y  =  "Percentage of Theses done in each Establishment in a Year " )
```



In []: