Exploring dataset "impreseFVG"

The dataset is organizes in a number of files; each file will be loaded in a different *data.frame*.

```
data.files <- list.files(path, pattern = ".csv$", recursive = TRUE)
print(paste("dataset contains",length(data.files), "files:"))
```

```
## [1] "dataset contains 10 files:"
```

```
print(data.files)
```

```
##  [1] "bilanci-fvg.csv"          "d_ateco.csv"
##  [3] "d_ng.csv"                 "id_imp_loc.csv"
##  [5] "pseudo_cf_id_impresa.csv" "t_attivita.csv"
##  [7] "t_codici.csv"             "t_imprese.csv"
##  [9] "t_imprese_dp.csv"         "t_localizz.csv"
```

imprese

The core data identifying companies can be found in *t_imprese.csv* .

```
imprese <- read.csv( paste0(path,"/t_imprese.csv"), sep = "|")
str(imprese)
```

```
## 'data.frame':    108379 obs. of  34 variables:
##  $ ï..fonte       : chr  "I" "I" "I" "I" ...
##  $ mm_aaaa        : chr  "06_2021" "06_2021" "06_2021" "06_2021" ...
##  $ id_impresa     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ denominazione  : chr  "PELLIZZARI SILVIO DI SEVERINO PELLIZZARI E C. S.N.C." "B.F.B.   CASA DI SPE
##  $ cf             : chr  "00000470310" "00002070324" "00002130938" "00003930328" ...
##  $ piva           : num  470310 2070324 2130938 3930328 4180931 ...
##  $ prov           : chr  "GO" "TS" "PN" "TS" ...
##  $ reg_imp_n      : chr  "GO007-1352" "TS006-7084" "PN033-2369" "TS006-4795" ...
##  $ sede_ul        : chr  "SEDE" "SEDE" "UL-1" "SEDE" ...
##  $ n.albo_art     : chr  "" "" "" "" ...
##  $ reg_imp_sez    : chr  "0" "0" "0" "0" ...
##  $ ng2            : chr  "SN" "SR" "SN" "AS" ...
##  $ stato_impresa  : chr  "INATTIVA" "ATTIVA" "INATTIVA" "ATTIVA" ...
##  $ data_cost      : chr  "1974-08-26" "1969-01-30" "1973-10-09" "1965-06-18" ...
##  $ data_isc_ri    : chr  "1996-02-19" "1996-02-19" "1996-02-19" "1996-02-19" ...
##  $ data_isc_rd    : chr  "1975-01-14" "1969-01-30" "1973-10-31" "1965-07-08" ...
##  $ data_isc_aa    : chr  "" "" "" "" ...
##  $ data_canc      : logi  NA NA NA NA NA NA ...
##  $ data_ini_at    : chr  "" "1969-01-30" "" "1965-06-18" ...
##  $ data_cess_att  : chr  "" "" "2008-05-21" "" ...
##  $ data_fall      : chr  "" "" "" "" ...
##  $ data_liquid    : chr  "" "" "" "" ...
##  $ addetti_aaaa   : int  1999 2015 0 2008 2009 2010 2013 1997 2015 0 ...
##  $ addetti_indip  : int  0 6 0 0 6 1 20 0 0 0 ...
##  $ addetti_dip    : int  0 39 0 2 2 0 24 0 80 0 ...
##  $ capitale       : num  NA 20000 0 0 0 ...
```

```
##  $ capitale_valuta: chr  "" "EURO" "EURO" "EURO" ...
##  $ imp_sedi_ee    : logi  NA NA NA NA NA NA ...
##  $ imp_eefvg      : chr  "" "" "" "" ...
##  $ imp_pmi        : chr  "NO" "NO" "NO" "NO" ...
##  $ imp_startup    : chr  "NO" "NO" "NO" "NO" ...
##  $ imp_femminile  : chr  "NO" "NO" "NO" "NO" ...
##  $ imp_giovanile  : chr  "NO" "NO" "NO" "NO" ...
##  $ imp_straniera  : chr  "NO" "NO" "NO" "NO" ...
```

The attributes belong to different groups:

- *metadata*:ï..fonte, mm_aaaa:
- *identifier*: id_impresa,reg_imp_n, cf, piva, denominazione
- *address*:prov,sede_ul,n.albo_art,reg_imp_sez

- *type of company*: ng2
- *active status*: stato_impresa

- *dates*:data_ini_at, data_cess_att, data_fall, data_liquid, data_cost, data_isc_ri, data_isc_rd,data_isc_aa,data_ca
- *employees*: addetti_aaaa, addetti_indip, addetti_dip

- *share capital*: capitale, capitale_valuta
- *other attributes*: imp_startup, imp_femminile, imp_giovanile, imp_straniera, imp_pmi, imp_sedi_ee, imp_eefvg

Metadata

Metadata are generated by the pre-rpcessing algorithm and provide information about source and last update. The two attributes (ï..fonte, mm_aaaa) are not relevant at this stage.

```
imprese <- subset(imprese, select = -c(ï..fonte, mm_aaaa))
```

Identifiers

The following attributes are relevant: - denominazione: company name - cf ("codice fiscale"): unique identifier, as factor (11 numbers or a string of 16 lettersa nd numbers) - id_impresa: unique identifier, numeric. Id and cf are unique, while company names are not and there are no missing values.

```
imprese$cf <- as.factor(imprese$cf)
imprese$denominazione <- as.factor(imprese$denominazione)
# check missing calues
sum(is.na(imprese$denominazione)) + sum(is.na(imprese$cf)) == 0
```

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

```
# check duplicates in cf
length(unique(imprese$cf)) == length(imprese$cf)
```

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

```
# check duplicates in denominazione
uniqueNames <-length(unique(imprese$denominazione))
allNames<-length(imprese$denominazione)
print(paste("Company names are not a valid identifier for further analysis: the dataset contains", uniq
```

```
## [1] Company names are not a valid identifier for further analysis: the dataset contains 107072 disti
```

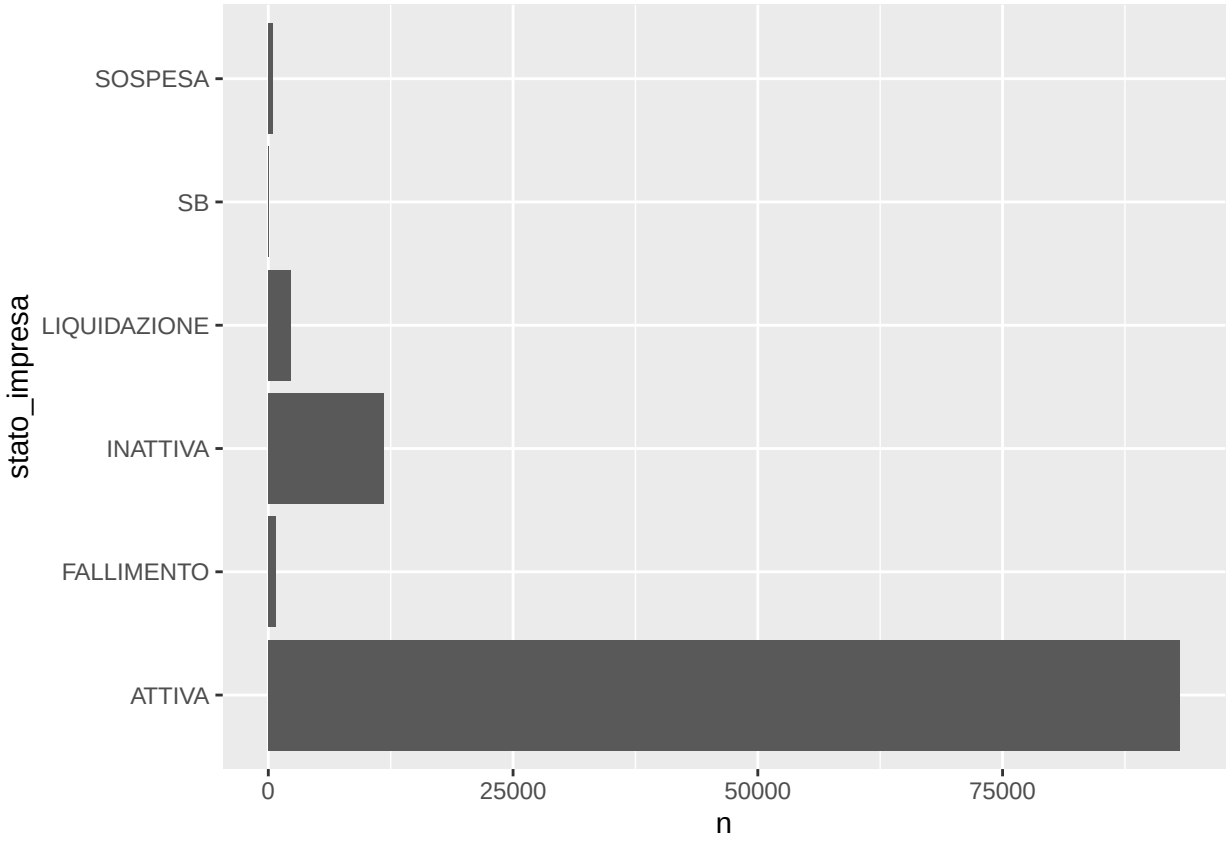
Other attributes (reg_imp_n,piva, n.albo_art,reg_imp_sez) are not relevant at this stage, and can be dropped.

```
imprese <- subset(imprese, select = -c(reg_imp_n,piva, n.albo_art,reg_imp_sez))
```

active status

Companies that are not active (e.g. due to bankruptcy, liquidation or suspended) are not relevant for the research objectives and can be removed from the dataset.

```
imprese$stato_impresa <- as.factor(imprese$stato_impresa)
df<-imprese %>% count(stato_impresa)
ggplot(data=df, aes(x=stato_impresa, y=n)) +   geom_bar(stat="identity") + coord_flip()
```



```
imprese<- subset(imprese, stato_impresa =='ATTIVA')
print(paste("Number of active companies: ", nrow(imprese)), quote=FALSE)
```

```
## [1] Number of active companies:  93106
```

location

TODO prov: province (GO, TS, UD, PN) » factor FVG / ITA / EU sede_ul: “SEDE” or “UL-n” » factor SEDE = HeadOffice / UL = LocalUnit LucalUnit = numeric 0 for HeadOffice, otherwise n To be transformed in factors

```
locs <- read.csv( paste0(path,"/t_localizz.csv"), sep = "|")
locs <- subset(locs, select = -c(ï..fonte))#ignore metadata
#connect to company id
str(locs)
```

```
## 'data.frame':    205385 obs. of  9 variables:
##  $ id_localiz         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ id_impresa         : int  1 2 2 2 3 4 5 6 7 8 ...
##  $ denominazione      : chr  "PELLIZZARI SILVIO DI SEVERINO PELLIZZARI E C. S.N.C." "B.F.B.   CASA DI
##  $ tipo_localizzazione: chr  "SE" "SE" "UL" "UL" ...
##  $ data_apert_ul      : chr  "" "" "2007-08-01 00:00:00" "2015-10-15 00:00:00" ...
##  $ prov_localiz       : chr  " GO" " TS" " TS" " TS" ...
##  $ comune             : chr  "CORMONS" "TRIESTE" "MONRUPINO" "TRIESTE" ...
##  $ indirizzo          : chr  "VIA PESCHERIA 4" "VIA CORTI 2" "FERNETTI 5" "PUNTO FRANCO NUOVO  EX CU
##  $ tipo_sedeul        : chr  "" "" "1:  > U - UFFICIO" "1:  > U - UFFICIO" ...
```

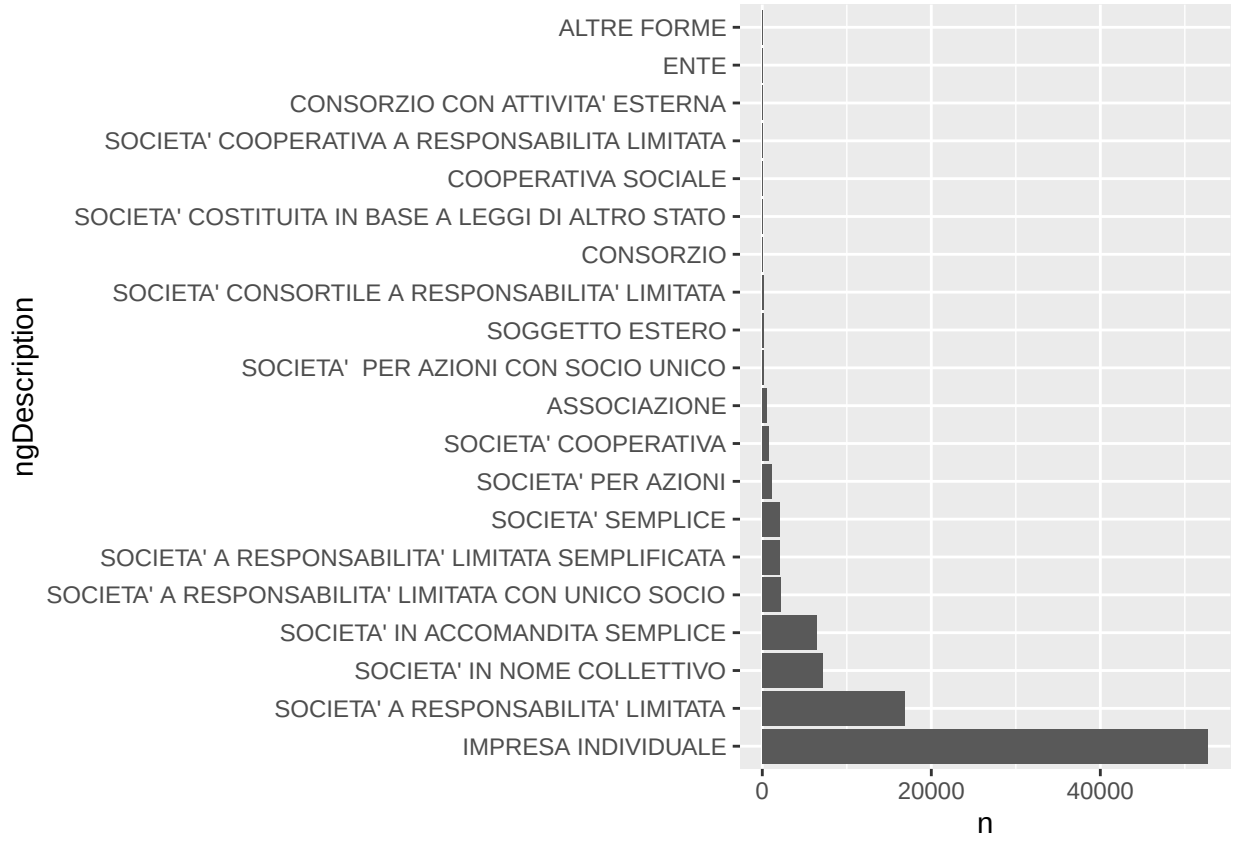
type of company

TODO: use ng2

```
# company type: keep only the relevan ones for the scope of our research.
types <- read.csv( paste0(path,"/d_ng.csv"), sep = "|")
imprese$ng2           <- as.factor(imprese$ng2)
names(types)<-c("ngGroup", "ng2", "ngDescription")

df <- imprese  %>% count(ng2)
df <- df %>% inner_join(types)
```

```
## Joining, by = "ng2"
```

```
df <- df  %>% arrange(-n) %>% head(20)
df$ngDescription <- factor(df$ngDescription, levels = df$ngDescription) #lock factors to keep the same
ggplot(data=df, aes(x=ngDescription, y=n)) +   geom_bar(stat="identity") + coord_flip()
```



Some company types are not relevant for our research, for example individual companies (DI) and other specified below. Dropping the corresponding dataframe rows drastically reduces the size of the data set

```
notRelevant = c("DI", "AZ", "IR", "ER", "EP", "EN", "EM", "EL", "EE", "SM", "MA", "SZ", "LL", "AM", "AF
toBeRemoved<-which(imprese$ng2 %in% notRelevant)
imprese2<-imprese[-toBeRemoved,]
print(nrow(imprese2))
```

```
## [1] 40211
```

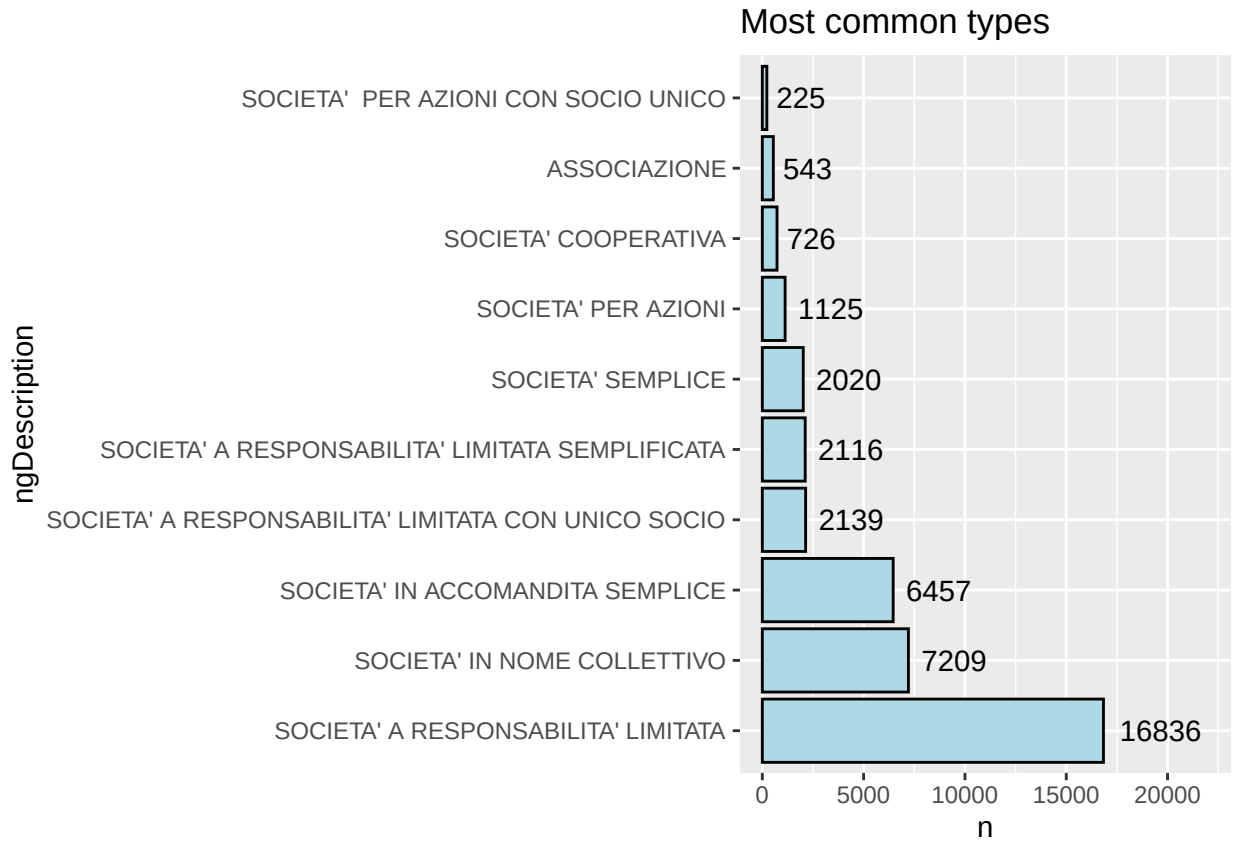
```
df <- imprese2  %>% count(ng2)
print(paste("The dataset contains ", nrow(df), "types of companies."), quote=FALSE)
```

```
## [1] The dataset contains  34 types of companies.
```

```
df <- df %>% inner_join(types)
```

```
## Joining, by = "ng2"
```

```
df <- df  %>% arrange(-n)
df$ngDescription <- factor(df$ngDescription, levels = df$ngDescription) #lock factors to keep the same
ggplot(data=head(df, 10), aes(x=ngDescription, y=n))  + geom_bar(stat="identity", color = "black", fill
```



dates, age of companies, years in business The dataset provides several dates: start of activity, dates of bankarupcy and cancellation We are interested in a broader information: "years in business"

Information on bankarupcy should be redundant, since non-active companies have been removed; nevertheless, past check showed unconistend data in the original tables. If bancarupcy dates are present, the company is considered as non-actie and removed.

at the end we keep only eta; all data fields can be removed

```
imprese$data_cost   <- as.Date(imprese$data_cost)
imprese$data_isc_aa <- as.Date(imprese$data_isc_aa)
imprese$data_isc_rd <- as.Date(imprese$data_isc_rd)
imprese$data_isc_ri <- as.Date(imprese$data_isc_ri)
imprese$data_ini_at <- as.Date(imprese$data_ini_at)

sum(is.na(imprese$data_isc_aa))
```

```
## [1] 63054
```

```
sum(is.na(imprese$data_isc_rd))
```

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

```
sum(is.na(imprese$data_isc_ri))
```

```
## [1] 1052
```

```r
sum(is.na(imprese$data_ini_at))
```

```
## [1] 1512
```

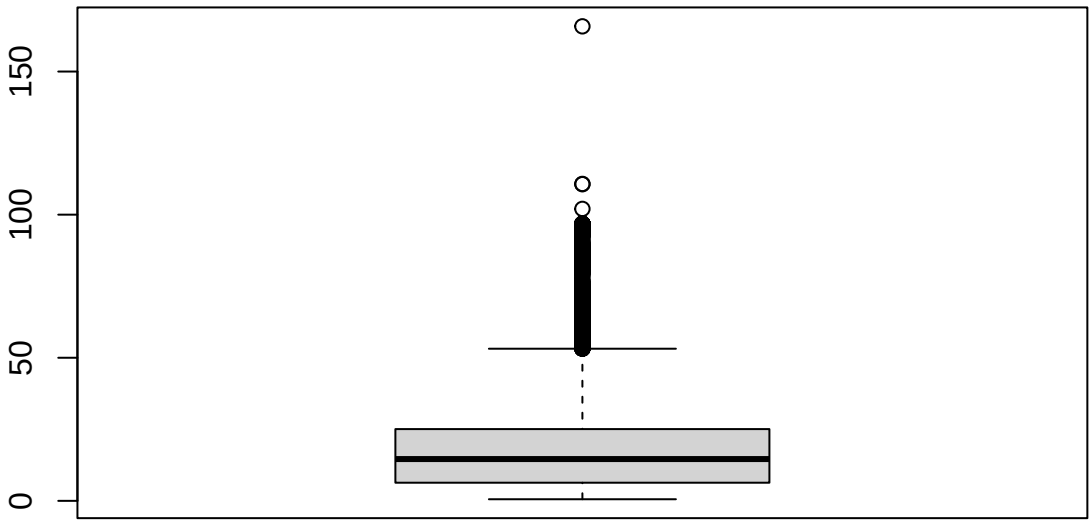
```r
#imprese %>% mutate(data_min = min(data_isc_aa, data_isc_rd, data_isc_ri, data_ini_at))
imprese["data_min"]<-imprese$data_isc_rd
sum(is.na(imprese$data_min))
```

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

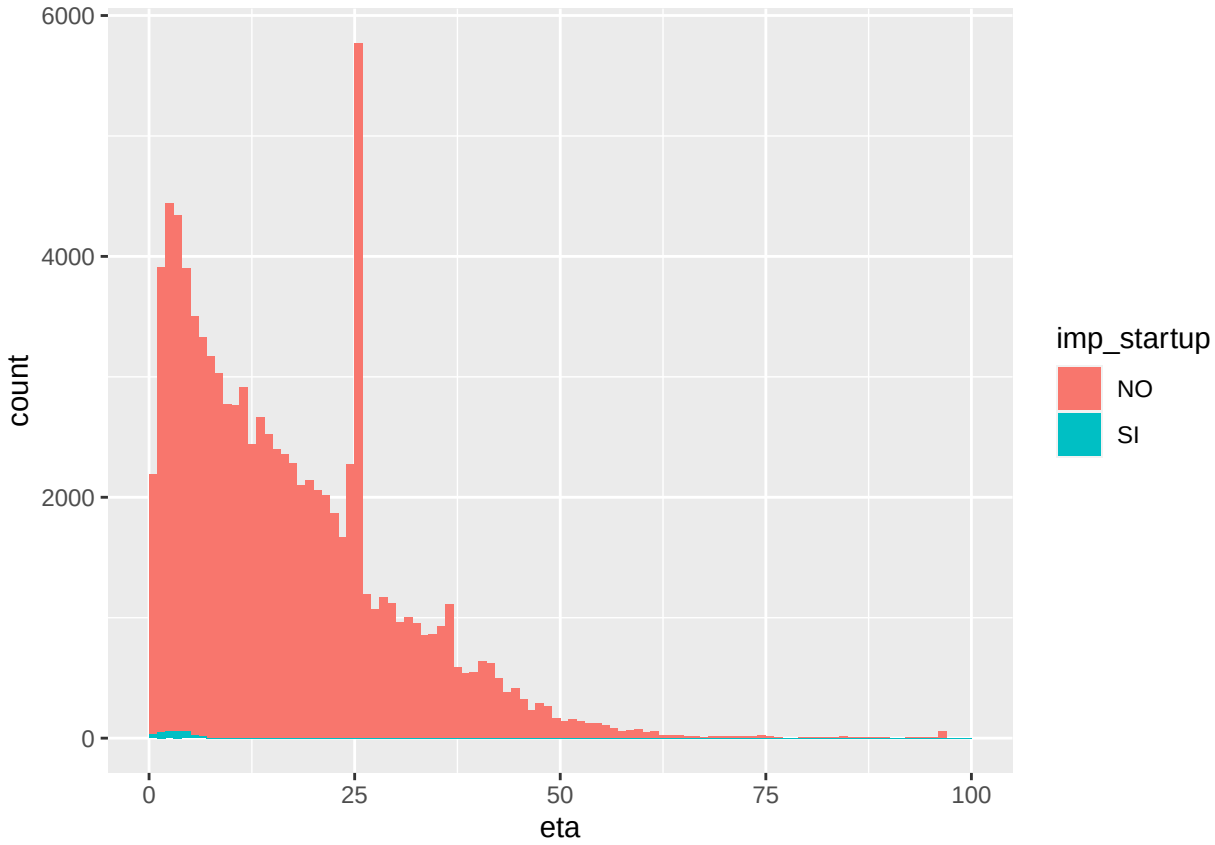
```r
imprese <- imprese[complete.cases(imprese$data_min), ]
imprese["eta"] = as.numeric(as.Date("2022-01-01") - imprese$data_min) / 365  #calcola età
imprese <- imprese[complete.cases(imprese$eta), ]
imprese <- imprese %>%
  select(-c(data_isc_aa, data_isc_rd, data_isc_ri, data_ini_at))

# TODO: select the minumum date
# %>%
#   filter(data_fall != "")     %>%                # drop unconsistent rows
#   filter(data_liquid != "")   %>%                # drop unconsistent rows
#   filter(data_cess_att != "") %>%                # drop unconsistent rows
#   filter(data_canc != "")                        # drop unconsistent rows
```

#TODO #imprese %>% mutate(data_min = coalesce(data_min, data_ini_at)) #verificare se funziona come previsto #imprese %>% mutate(data_min = coalesce(data_min,data_isc_aa)) #imprese %>% mutate(data_min = coalesce(data_min,data_isc_rd)) #imprese %>% mutate(data_min = coalesce(data_min,data_isc_ri))

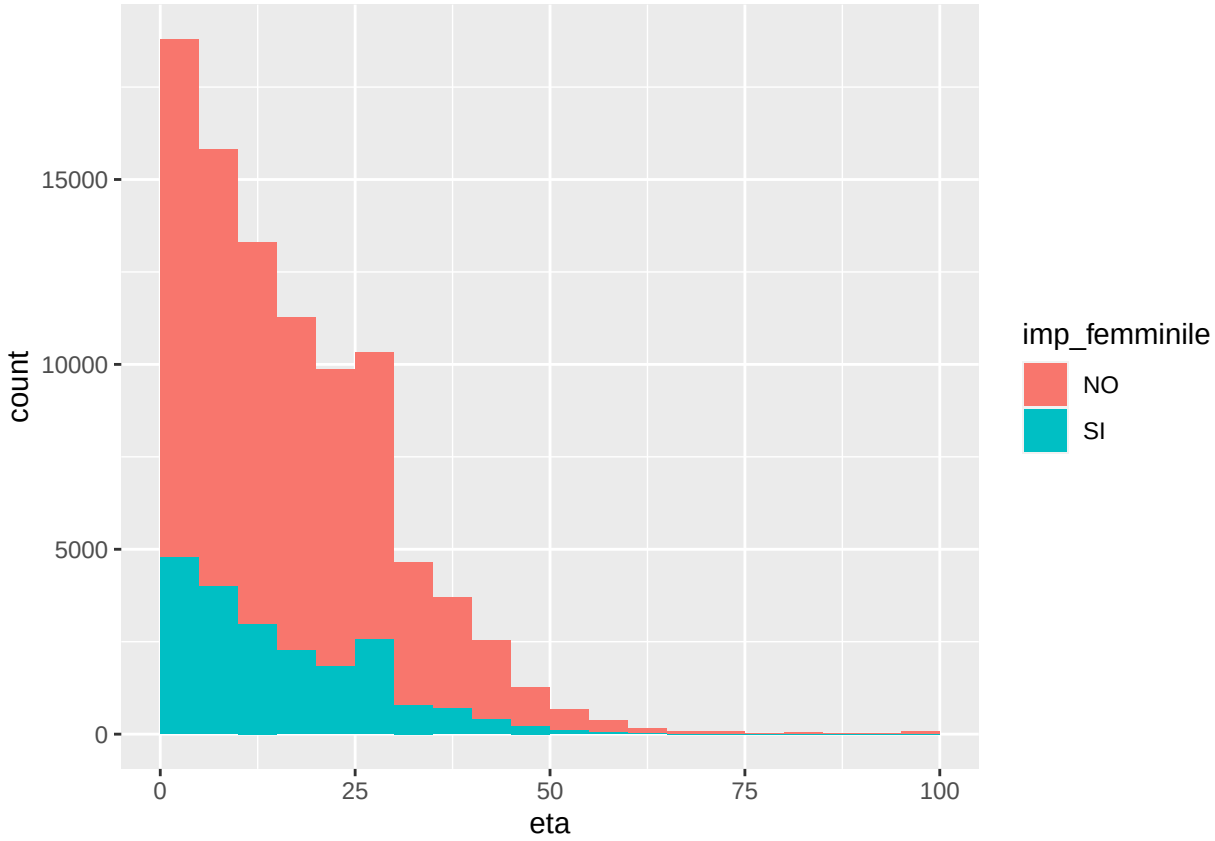
```r
eta <- imprese$eta
boxplot(imprese$eta)
```




```
ggplot(data=imprese, aes(x=eta, fill = imp_startup)) + geom_histogram(breaks=seq(0, 100, by=1))
```




```
ggplot(data=imprese, aes(x=eta, fill = imp_femminile)) + geom_histogram(breaks=seq(0, 100, by=5))
```



TODO use all "start dates" to get a better estimate of *years in business*

*:data_ini_at, data_cess_att, data_fall, data_liquid, data_cost, data_isc_ri, data_isc_rd,data_isc_aa,data_canc

employees

addetti_aaaa, addetti_indip, addetti_dip

share capital

capitale, capitale_valuta

other attributes

TODO other attributes are relevant as signs of innovation imp_startup, imp_femminile, imp_giovanile, imp_straniera, imp_pmi, imp_sedi_ee, imp_eefvg
show distribution for each

```
imprese$imp_pmi       <- as.factor(imprese$imp_pmi)
imprese$imp_startup   <- as.factor(imprese$imp_startup)
imprese$imp_giovanile <- as.factor(imprese$imp_giovanile)
imprese$imp_femminile <- as.factor(imprese$imp_femminile)
imprese$imp_straniera <- as.factor(imprese$imp_straniera)
imprese$imp_sedi_ee   <- as.factor(imprese$imp_sedi_ee)
```

Codici

Each company is associated with a list of activity codes, according to NACE classification. The corresponding information is available in *t_codici.csv* . As in previous files, metadata can be inored at this stage. More info: Complete list of all NACE Code NACE (Nomenclature of Economic Activities) is the European statistical classification of economic activities. NACE groups organizations according to their business activities. [https://nacev2.com/en]

> TODO nace codes are associated with a company and each of its localizations. We may be interesed in reducin the compexity, summarizing all codes, and ignoring association to each location. But this can be an issue in some cases. For example a company has its head office in Rome, with some codes, and a local unit in Trieste engaged in different activities. Shall we consider all the codes, or by location? TODO Some companies have no nace codes associated: how many? should we drop the corresponding row in imprese? TODO Explore code types (I P S) they are not unique.

```
naceCodes <- read.csv( paste0(path,"/t_codici.csv"), sep = "|")  %>%
  select(-c(ï..fonte, mm_aaaa)) %>%         #ignore metadata
  filter(ateco != "") %>%                 # drop rows with empty nace codes
  inner_join(locs, by="id_localiz") %>%
  select(-c(data_apert_ul, comune, indirizzo, tipo_sedeul))

naceCodes["sector"]<-substr(naceCodes$ateco,start=1,stop=2)
str(naceCodes)
```

```
## 'data.frame':    402374 obs. of  9 variables:
##  $ id_localiz         : int  2 2 2 3 3 4 4 6 6 7 ...
##  $ loc_n              : int  0 0 0 2 2 4 4 0 0 0 ...
##  $ ateco_tipo         : chr  "I" "P" "S" "P" ...
##  $ ateco              : chr  "52.29.1" "52.29.1" "49.41" "52.29.1" ...
##  $ id_impresa         : int  2 2 2 2 2 2 2 4 4 5 ...
##  $ denominazione      : chr  "B.F.B.   CASA DI SPEDIZIONI S.R.L." "B.F.B.   CASA DI SPEDIZIONI S.R.L
##  $ tipo_localizzazione: chr  "SE" "SE" "SE" "UL" ...
##  $ prov_localiz       : chr  " TS" " TS" " TS" " TS" ...
##  $ sector             : chr  "52" "52" "49" "52" ...
```

Not all sectors are relevant for the research topic. We can restrict the dataset to manufacturing activities (codes 10 to 33)and research (sector = 72)

```
relevantSectors <- seq(10,33)
relevantSectors <- c(relevantSectors,72)
relevantSectors
```

```
##  [1] 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 72
```

```
naceCodes <- naceCodes %>% filter(sector %in% relevantSectors)      #filtering codes
imprese <- imprese %>% filter(id_impresa %in% naceCodes$id_impresa) #filtering companies

print(paste("Number of companies reduced to ", nrow(imprese)))
```

```
## [1] "Number of companies reduced to  12918"
```