Exploring dataset "companiesFVG"

This section is deditated to the exploration of data on companies that have a localization in Friuli Venezia Giulia. The origin of data is the Italian Business registry (https://www.registroimprese.it/il-registro-imprese-per-la-p.a.), and is managed by Infocamere (https://www.infocamere.it/).

```
data.files <- list.files(pathRawData, pattern = ".csv$", recursive = TRUE)
```

Pre-processed is availavle in folder **/data/raw**, organized in 10 files (namely bilanci-fvg.csv, d__ateco.csv, d__ng.csv, id__imp__loc.csv, pseudo__cf__id__impresa.csv, t__attivita.csv, t__codici.csv, t__imprese.csv, t__imprese__dp.csv, t__localizz.csv). Each will be loaded in a different *data.frame* and, after feature selection, saved in a new folder **/data/tidy**.

companies

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

```
companies <- read_delim( paste0(pathRawData,"/t_imprese.csv"))
```

```
## Rows: 108379 Columns: 34
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "|"
## chr  (19): fonte, mm_aaaa, denominazione, cf, piva, prov, reg_imp_n, sede_ul...
## dbl   (5): id_impresa, addetti_aaaa, addetti_indip, addetti_dip, capitale
## lgl   (2): data_canc, imp_sedi_ee
## date  (8): data_cost, data_isc_ri, data_isc_rd, data_isc_aa, data_ini_at, da...
```

```
##
## 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.
```

```
spec(companies) # tydiverse for str(companies)
```

```
## cols(
##   fonte = col_character(),
##   mm_aaaa = col_character(),
##   id_impresa = col_double(),
##   denominazione = col_character(),
##   cf = col_character(),
##   piva = col_character(),
##   prov = col_character(),
##   reg_imp_n = col_character(),
##   sede_ul = col_character(),
##   ‘n-albo_art‘ = col_character(),
##   reg_imp_sez = col_character(),
##   ng2 = col_character(),
##   stato_impresa = col_character(),
##   data_cost = col_date(format = ""),
##   data_isc_ri = col_date(format = ""),
##   data_isc_rd = col_date(format = ""),
##   data_isc_aa = col_date(format = ""),
##   data_canc = col_logical(),
##   data_ini_at = col_date(format = ""),
##   data_cess_att = col_date(format = ""),
##   data_fall = col_date(format = ""),
##   data_liquid = col_date(format = ""),
##   addetti_aaaa = col_double(),
##   addetti_indip = col_double(),
##   addetti_dip = col_double(),
##   capitale = col_double(),
##   capitale_valuta = col_character(),
##   imp_sedi_ee = col_logical(),
##   imp_eefvg = col_character(),
##   imp_pmi = col_character(),
##   imp_startup = col_character(),
##   imp_femminile = col_character(),
##   imp_giovanile = col_character(),
##   imp_straniera = col_character()
## )
```

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. Current version is 06_2021. The two attributes (ï..fonte, mm_aaaa) are not relevant for further analysis, and can be removed from the tidy dataset.

```
companies <- companies %>% select( !c(fonte, mm_aaaa))
```

Identifiers

Each company can be identified by its name, vat number, fiscal code, The following attributes will be used in: - denominazione: company name (not a unique identifier, can be spelled in different ways across datasets; moreover different companies may have the same name); - cf ("codice fiscale"): unique identifier, as factor. Values are unique for each company, but the structure depends on company type: generally a string of 11 digits, but individual companies refer to the owner's code, a string of 16 letters and digits; -

idCompany ("id_impresa"): unique identifier, numeric, created in pre-processing phase. Other attributes (reg_imp_n,piva, n.albo_art,reg_imp_sez) are not relevant at this stage, and can be dropped. All identifiers will be converted to factors.

```
companies <- companies %>%
             select( !c(reg_imp_n, piva, `n-albo_art`,reg_imp_sez)) %>%
             rename(name = denominazione) %>%
             rename(idCompany = id_impresa) %>%
             mutate_if(is.character, as.factor)
```

Now we can ckeck if there are any missing values or duplicates

```
# check missing calues
sum(is.na(companies$name)) + sum(is.na(companies$cf)) == 0
```

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

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

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

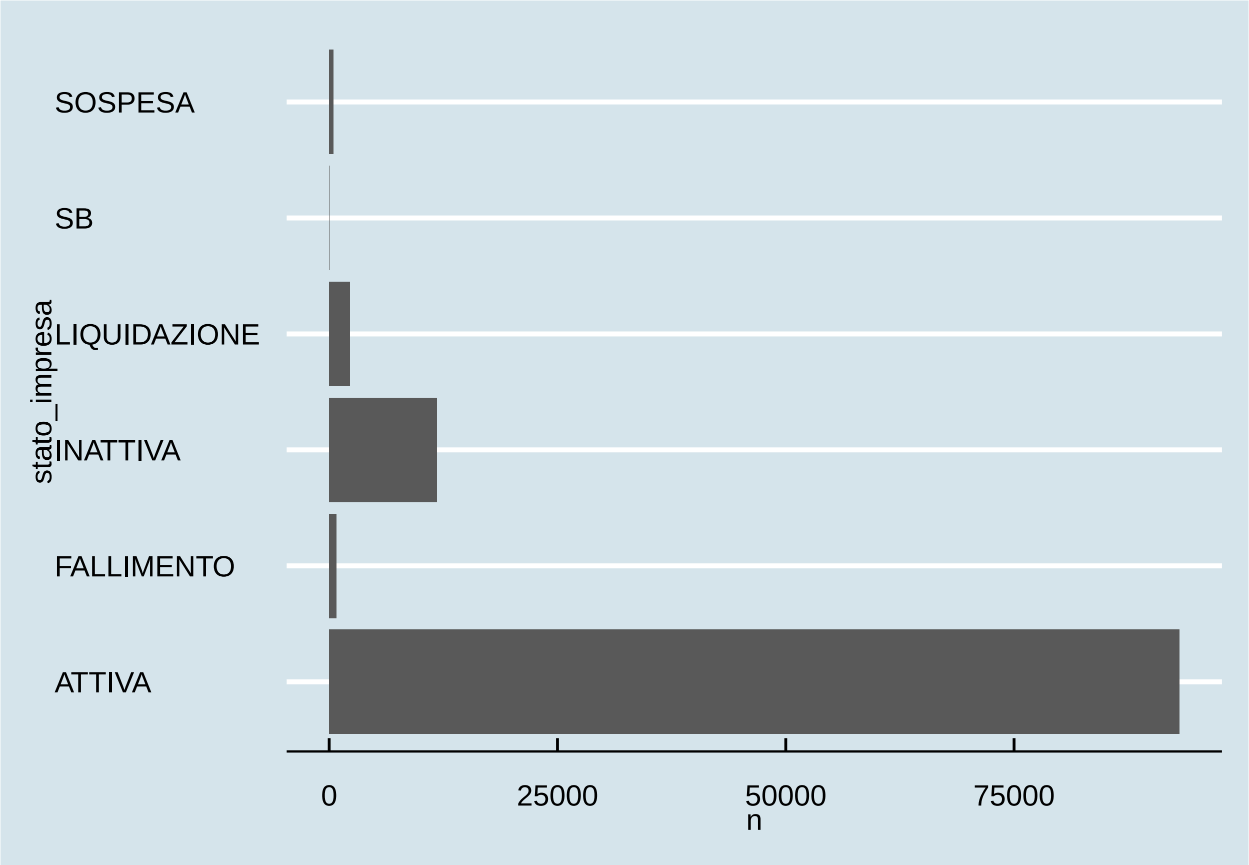
```
# check duplicates in name
uniqueNames <-length(unique(companies$name))
allNames<-length(companies$name)
print(paste("Company names are not a valid identifier for further analysis: the dataset contains", uniq
```

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

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.

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



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

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

head office and local units

Each company has a "registered office" (sede legale) and may have several local units (unità locale). Relevant data is stored in file "t_localizz.csv". > TODO: add description of variables. Select only companies that are located in Friuli Venezia Giulia, according to prov: province (GO, TS, UD, PN). Select companies that have head office abroad. sede_ul: "SEDE" or "UL-n" » factor SEDE = HeadOffice / UL = LocalUnit Extract the number of local units from attribute "sede_ul" Create new variable "head-office" true/false use "data apertura ul" to improve the estimate of company age (or remove the attribute) remove unnecessary variables

```
locs <- read_delim( paste0(pathRawData,"/t_localizz.csv")) %>%
             select( c(id_localiz, id_impresa, denominazione,tipo_localizzazione)) %>%
             rename(name = denominazione) %>%
             rename(idCompany = id_impresa) %>%
             mutate_if(is.character, as.factor)
```

```
## Rows: 205385 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "|"
## chr  (7): fonte, denominazione, tipo_localizzazione, prov_localiz, comune, i...
## dbl  (2): id_localiz, id_impresa
## dttm (1): data_apert_ul
```

```
##
## 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.
```

Legal form of companies

The legal form of companies can be a relevant attribute for further research. A primary distinction should be made between between

- *limited liability companies* (società di capitali), mainly Private Limited Companies by Quotas (società a responsabilità limitata or S.r.l.), Simplified S.r.l. (S.r.l.s.), Public Limited Companies by Shares (società per azioni or S.p.A.). Limited liability companies disclose their financial statements, therefore more relevant information is available.
- *partnerships* (società di persone), mainly Società in nome collettivo (S.n.c.) and Società in accomandita semplice or (S.a.s.)

Legal form is encoded in the variable companies$ng, and codes are described in a separate file */d_ng.csv*. There are over 50 different codes, but the great majority of companies belong to a small number of types. For the purpose of data exploration, the following figure hignliths thee 10 most common legal forms.

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

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

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

```
df <- df  %>% arrange(-n) %>% head(10)
df$ngDescription <- factor(df$ngDescription, levels = df$ngDescription)
df
```

```
## # A tibble: 10 x 4
##    ng2       n ngGroup              ngDescription
##    <chr> <int> <chr>                <fct>
##  1 DI    52751 IMPRESE INDIVIDUALI  IMPRESA INDIVIDUALE
##  2 SR    16836 SOCIETA' DI CAPITALE SOCIETA' A RESPONSABILITA' LIMITATA
##  3 SN     7209 SOCIETA' DI PERSONE  SOCIETA' IN NOME COLLETTIVO
##  4 AS     6457 SOCIETA' DI PERSONE  SOCIETA' IN ACCOMANDITA SEMPLICE
##  5 SU     2139 SOCIETA' DI CAPITALE SOCIETA' A RESPONSABILITA' LIMITATA CON UNI~
##  6 RS     2116 SOCIETA' DI CAPITALE SOCIETA' A RESPONSABILITA' LIMITATA SEMPLIF~
##  7 SE     2020 SOCIETA' DI PERSONE  SOCIETA' SEMPLICE
##  8 SP     1125 SOCIETA' DI CAPITALE SOCIETA' PER AZIONI
##  9 SC      726 ALTRE FORME          SOCIETA' COOPERATIVA
## 10 AC      543 ALTRE FORME          ASSOCIAZIONE
```

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

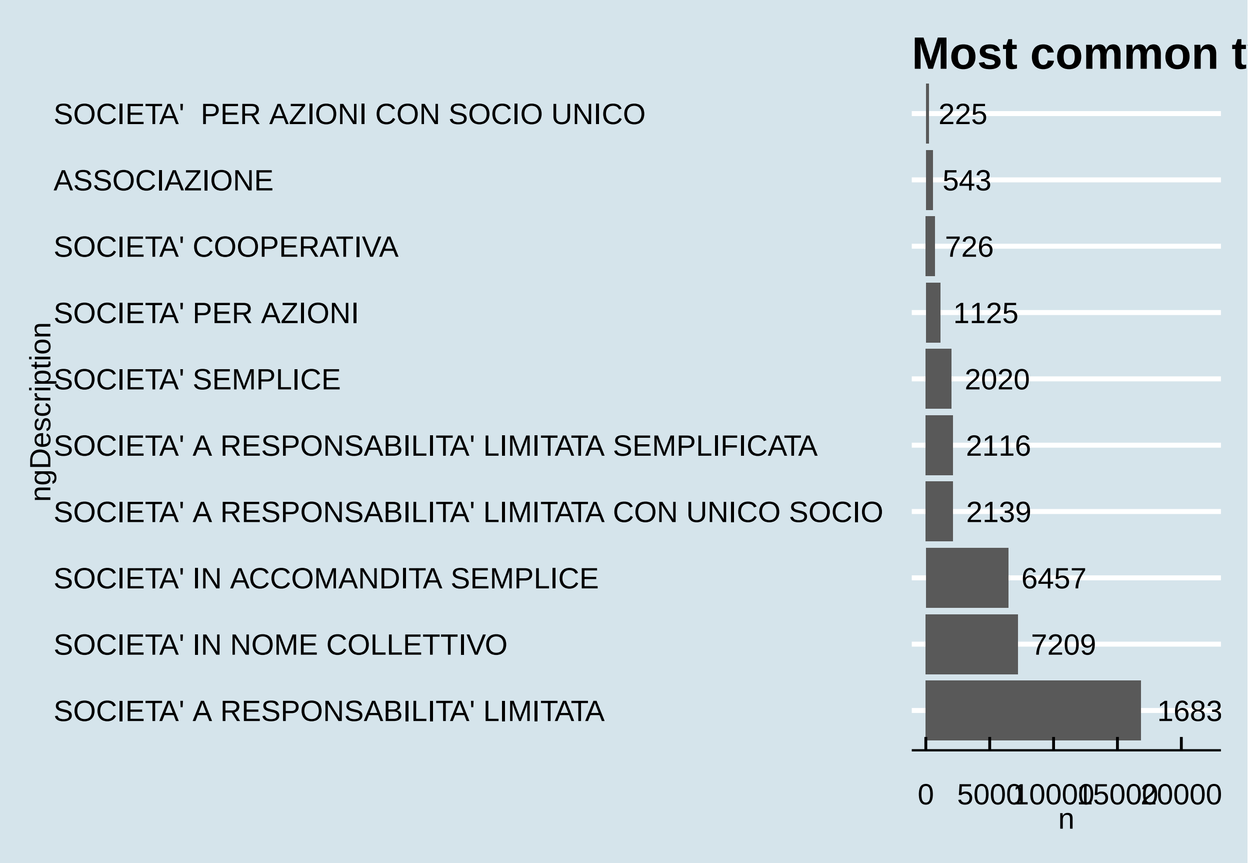
```
notRelevant = c("DI", "AZ", "IR", "ER", "EP", "EN", "EM", "EL", "EE", "SM", "MA", "SZ", "LL", "AM", "AF
toBeRemoved<-which(companies$ng2 %in% notRelevant)
companies<-companies[-toBeRemoved,]

df <- companies  %>% 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") + coord_flip() + ggti
```



dates, age of companies, years in business

The dataset provides relevant information in the form of dates.

```
companies %>% select(starts_with("data")) %>% summary(is.na())
```

```
##    data_cost           data_isc_ri          data_isc_rd        
##  Min.   :1807-01-01   Min.   :1996-02-08   Min.   :1856-04-26  
##  1st Qu.:1992-12-31   1st Qu.:1996-02-19   1st Qu.:1994-03-28  
##  Median :2005-01-11   Median :2005-11-07   Median :2005-10-20  
##  Mean   :2002-02-24   Mean   :2006-05-16   Mean   :2002-12-06  
##  3rd Qu.:2014-03-25   3rd Qu.:2014-12-01   3rd Qu.:2014-10-31  
##  Max.   :2021-05-28   Max.   :2021-06-01   Max.   :2021-05-31  
##  NA's   :1229         NA's   :805          NA's   :5           
##   data_isc_aa         data_canc       data_ini_at         data_cess_att  
##  Min.   :1937-10-24   Mode:logical   Min.   :1856-04-26   Min.   :NA     
##  1st Qu.:1989-07-06   NA's:40211     1st Qu.:1994-02-01   1st Qu.:NA     
##  Median :2002-04-19                  Median :2006-01-04   Median :NA     
##  Mean   :2000-01-31                  Mean   :2003-01-08   Mean   :NA     
##  3rd Qu.:2012-02-01                  3rd Qu.:2014-09-20   3rd Qu.:NA     
##  Max.   :2021-06-04                  Max.   :2105-02-28   Max.   :NA     
##  NA's   :32392                       NA's   :1041         NA's   :40211  
##    data_fall           data_liquid        
##  Min.   :1982-10-01   Min.   :2009-06-15  
##  1st Qu.:1997-07-03   1st Qu.:2010-12-25  
##  Median :2004-02-02   Median :2012-10-08  
##  Mean   :2003-10-16   Mean   :2013-11-12  
##  3rd Qu.:2008-11-05   3rd Qu.:2014-09-30  
##  Max.   :2021-05-18   Max.   :2019-12-20  
##  NA's   :39671        NA's   :40194       
```

The most relevant information for further research is the *age* of the company, that can be assessed as the number of years in business, i.e. the number of years from the earliest date of registration. There are several registration ates (data_iscr_*) and an official start date (data_ini_at), however each attribute has several missing values (NA).

```
#create new atribute "years in business"
companies <- companies %>%
  rowwise() %>%
  mutate(dateMin = min(data_isc_ri, data_isc_rd, data_isc_aa, data_ini_at,  na.rm=TRUE)) %>%
  mutate(yearsInBusiness = as.numeric(as.Date("2022-01-01") - dateMin) / 365  )
```

The only attribute to keep for further analysis is *yearsInBusiness*; all *date* columns can be removed from the tidy dataset.

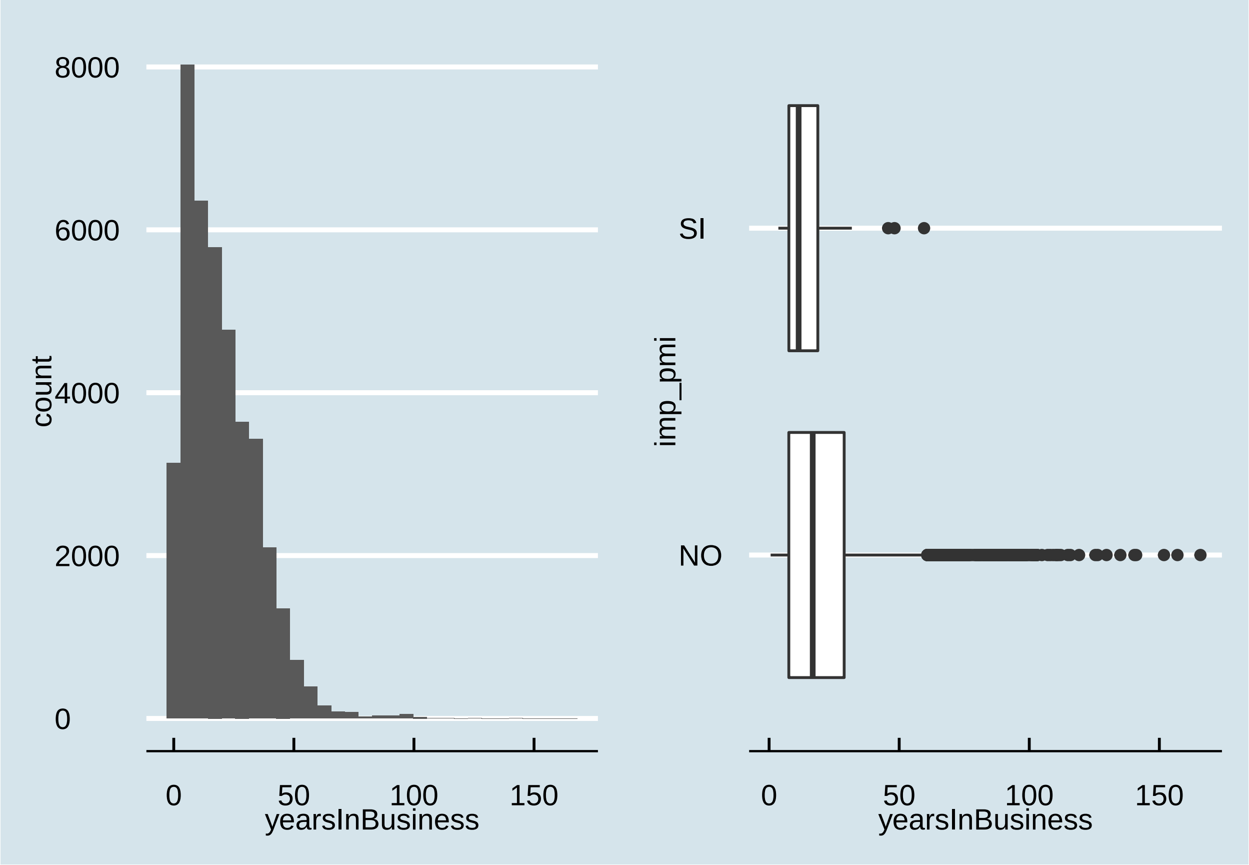
```
#remove all "date" sttributes
companies <- companies %>% select( !starts_with("data"))
```

The distribution of yearsInBusiness is shown below. A

```
plot1<-ggplot(data=companies, aes(x=yearsInBusiness), show.legend = FALSE) +
        geom_histogram()
plot2<-ggplot(data=companies, mapping=aes(x=imp_pmi, y=yearsInBusiness))+
        geom_boxplot()  + coord_flip()

labels = c("years in business", "focus on SMEs")
figure <- ggarrange(plot1, plot2)#,labels = labels, ncol = 1, nrow = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```



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.

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

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

NACE activity codes

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]

```
NACECodes <- read_delim( paste0(pathRawData,"/t_codici.csv"))  %>%
  select(-c(fonte, mm_aaaa))
```

```
## Rows: 409713 Columns: 6
```

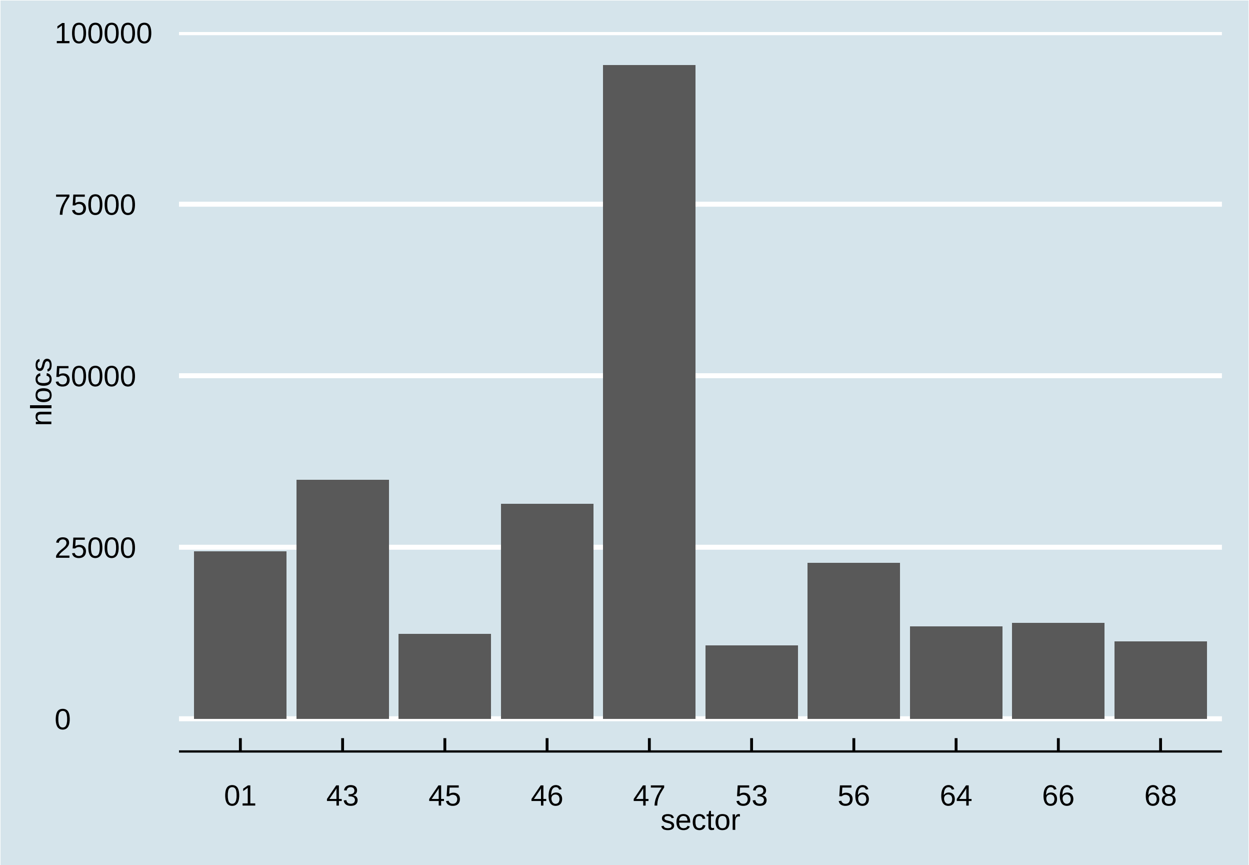
```
## -- Column specification --------------------------------------------------------
## Delimiter: "|"
## chr (4): fonte, mm_aaaa, ateco_tipo, ateco
## dbl (2): id_localiz, loc_n
```

```
##
## 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.
```

```
noCode <- NACECodes %>% filter(is.na(ateco))
percNA <- round(nrow(noCode)/nrow(NACECodes)*100, 2)
compNoCode <- noCode %>% distinct()
percCompNA <- round(nrow(compNoCode)/nrow(companies)*100,2)
```

The dataset contains 7339 companies, and 40211 NACE codes: each company may have one or more NACE codes, of different types: I (prevealente), P (Primario) and S (Secondario). Unicity is not guaranteed for any type (despite “I” codes are supposed to be unique, in fact many companies have several codes).

```
# TODO most common sectors (a sector is identified by first two digits of nacecode)
NACECodes["sector"]<-substr(NACECodes$ateco,start=1,stop=2)
df <- NACECodes %>% group_by(sector) %>% summarise( nlocs = n())  %>% arrange(desc(nlocs))   %>% head(1
p<-ggplot(data=df, aes(x=sector, y=nlocs)) +  geom_bar(stat="identity")
p
```



A company can be **associated with a single NACE code** trough an algorithm, than is to some extent subjective.

Some companies have **no NACE codes associated**: this is an issue with the original data affecting a small portion of the dataset (in this example 7339 missing values out of 409713, or 1.79%). Since each company may have more codes, the number of companies without *any NACE code* is smaller: only 4 out of 40211 i.e. 18.09%. Rows with NAs are removed from the dataset.

Moreover, NACE codes are **associated with company localization**. In order to reduce the compexity, we may summarize NACE codes by company, but this can lead to misinterpretation in some cases. Consider, for example, a company has its head office in Rome, with several codes in sector 47, and a local unit in Trieste engaged in different activities with a single code in sector 22. If we summarise codes by company our case study belongs to sector 47; instead, if we summarize by localization, the company belongs to sector 22.

For the purpose of data elxploration and preliminary feature selection, we filer companies and locs to retain only rows that are associated with NACE codes.

```
loc_in_NACE <- NACECodes %>% select(id_localiz) %>% distinct()    #unique id_localiz associated with NAC
locs <- locs %>% filter(id_localiz %in% loc_in_NACE$id_localiz)  #filter locs dataframe
companies <- companies %>% filter(idCompany %in% locs$idCompany) #filter companies dataframe
```

TODO sectors are identified by first two digits of NACE codes.

companies<- subset(companies2, select = c(‘eta’, ‘cf’, ‘id_impresa’)) ids <- unique(companies2$id_impresa)localiz < -read.csv(”t_localizz.csv”, sep = ”|”)localiz < -localiz[localiz$id_impresa %in% ids,] localiz <- subset(localiz, select = c(‘id_impresa’, ‘id_localiz’)) locs <- unique(localiz$id_localiz)codici < -read.csv(”t_codici.csv”, sep = ”|”)codici < -codici[codici$id_localiz %in% locs,] codici <- subset(codici, select = c(‘id_localiz’, ‘ateco’)) codici = merge(codici, localiz, by = “id_localiz”)

imp <- data.frame(“company”=ids) imp$eta < -companies$eta imp$ateco <- NA

” ’ ’ ’ ’”

tidy dataset

Now we can save the filtered and cleaned data to a csv. The number of features is reduced to… The dataset is composed of .. .main files (t_cmp.csv, t_nace.csv) and … files with extended descriptions (d_ng, d_nace).

```
companies %>% write_csv(paste0(pathTidyData,"t_cmp.csv"),)
```