Exploration of company information

Fabio Morea

2022-01-26

Contents

Chapter 1

Scope and objectives

Placeholder

1.1 Background information:

1.2 About this notebook

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"
```

1.3 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.I
##  $ 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  "O" "O" "O" "O" ...
##  $ 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:

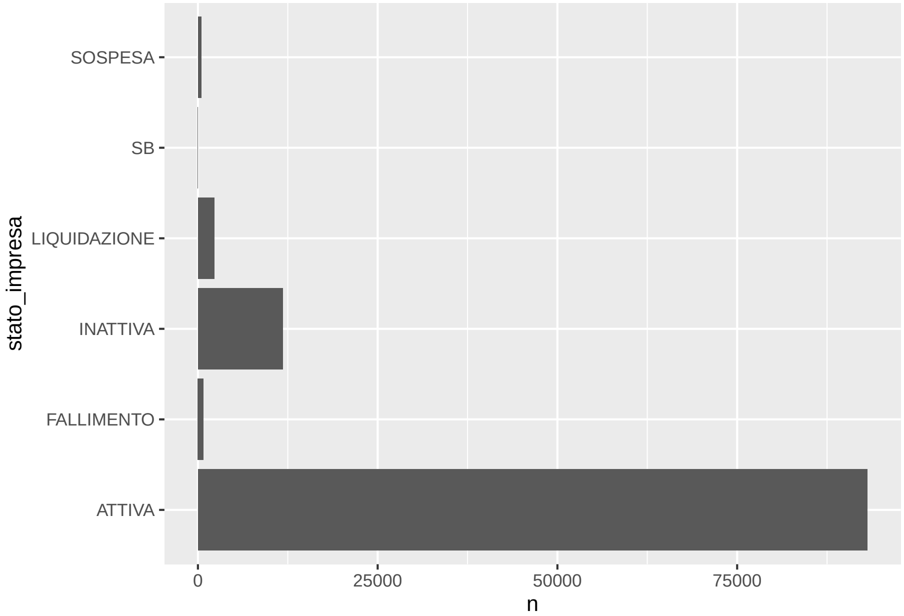
- *metadata*:ï..fonte, mm_aaaa:
- *identifier*: id_impresa,reg_imp_n, cf, piva, denominazione
- *address*:prov,sede_ul,n.albo_art,reg_imp_sez

- *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_canc
- *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

```
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')
nrow(imprese)
```

```
## [1] 93106
```

1.4 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.

1.5 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.

```r
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
```

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

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

```r
# 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 datase
```

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

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

1.6 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

- *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_canc
- *employees*: addetti_aaaa, addetti_indip, addetti_dip

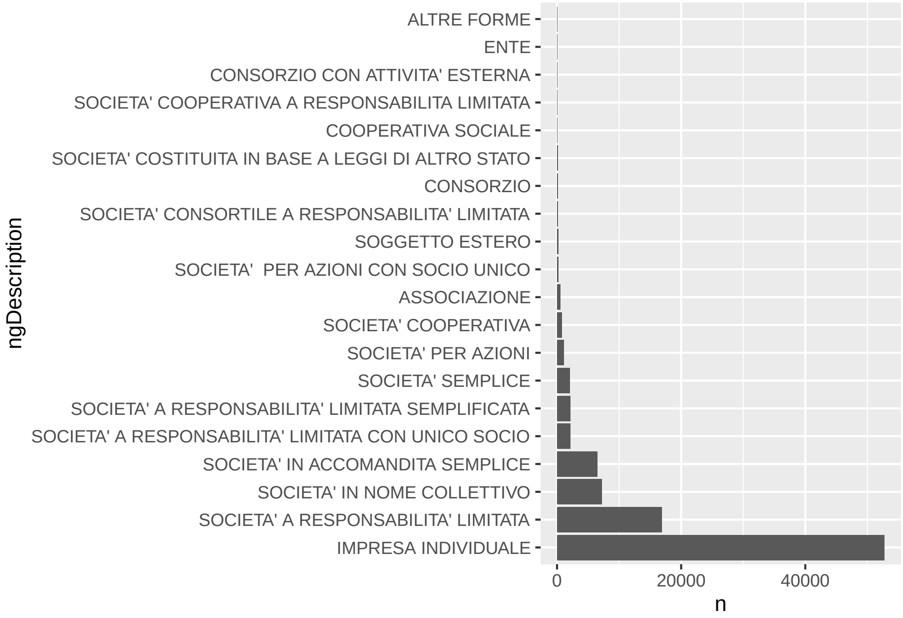
- *share capital*: capitale, capitale_valuta
- *other attributes*: imp_startup, imp_femminile, imp_giovanile, imp_straniera, imp_pmi, imp_sedi_ee, imp_eefvg

```
# 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
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
toBeRemoved<-which(imprese$ng2 %in% notRelevant)
```

```
imprese2<-imprese[-toBeRemoved,]
print(nrow(imprese2))
```

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

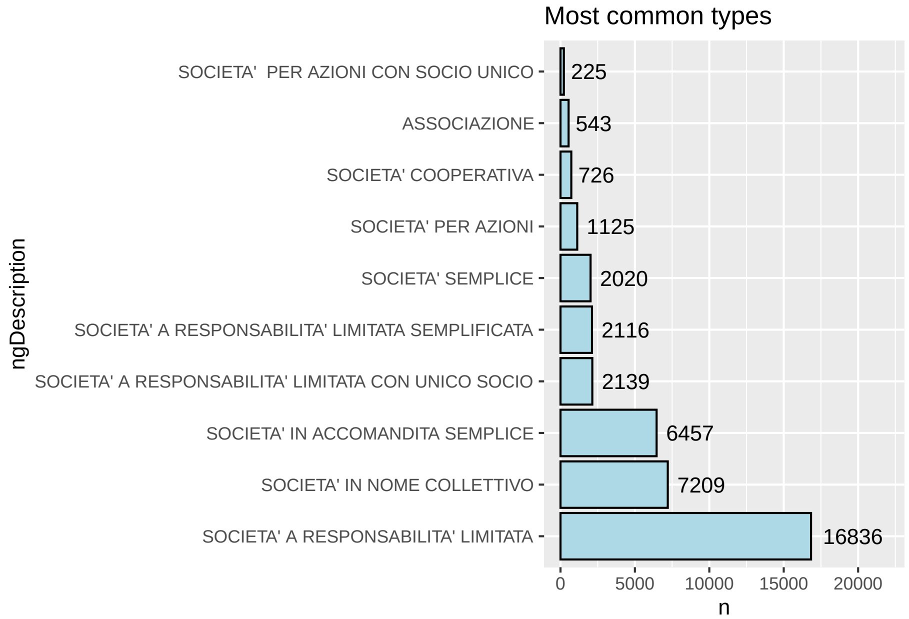
```
df <- imprese2  %>% count(ng2)
print(paste("The dataset contains ", nrow(df), "types of companies."), quote=FALSE)
```

```
## [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
ggplot(data=head(df, 10), aes(x=ngDescription, y=n))  + geom_bar(stat="identity", colo
```



Exploring dataset “bilanciFVG”

Placeholder

Exploring dataset “ratingFVG”

Placeholder

Exploring dataset “CO-FVG”

Placeholder

introduction

Placeholder

similarities based on metric features

custimized distances

Further development

Placeholder

1.7 distance

1.8 A similarity function between companies based on NACE codes