Exploring dataset CO-FVG

The original data is organized in 8 files: dati_2014.csv, dati_2015.csv, dati_2016.csv, dati_2017.csv, dati_2018.csv, dati_2019.csv, dati_2020.csv, dati_2021.csv. > TODO Currently, data exploration phase is focused on only one of the files above. Should extend it to all files using a for loop and appending results to a data.frame.

```
empl <- read_delim( paste0(pathRawData,"dati_2018.csv"))
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 395456 Columns: 43
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "|"
## chr  (25): CF, az_ragione_soc, genere, id_cittadino, professione, qualifica,...
## dbl   (8): ...1, anno, eta, mese, saldo, codice_istat, SLL_codice, qualifica...
## lgl   (5): somm, erroriEta, errori_qualifica, erroriCF, errori
## date  (5): data, data_fine, data_fine_prev, data_inizio, data_nascita
```

```
##
## 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.
```

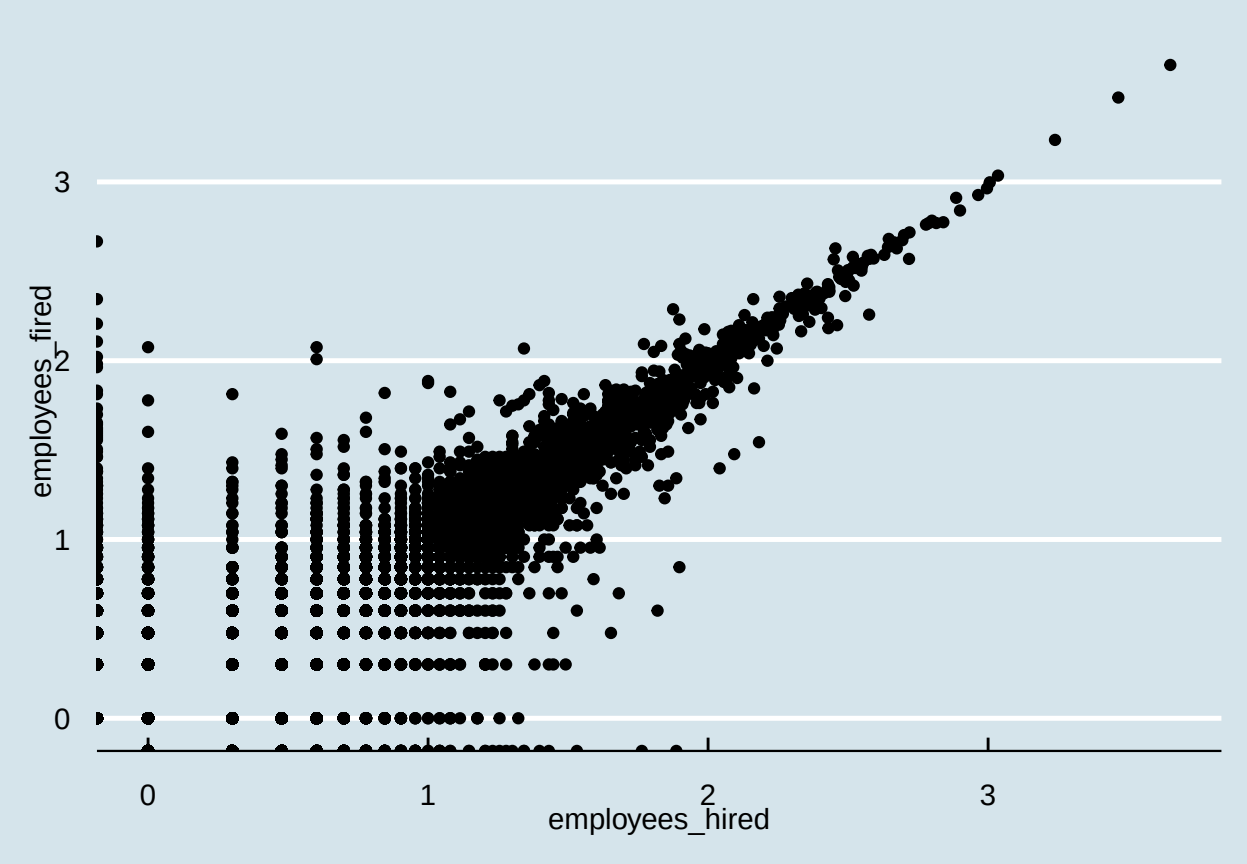
```
features <-names(empl)
some_features <- c("CF","anno","eta","genere","iso3","professione","qualifica","saldo")
```

There are 43 features available: ...1, CF, anno, az_ragione_soc, data, data_fine, data_fine_prev, data_inizio, data_nascita, eta, genere, id_cittadino, mese, professione, qualifica, qualifica_codice, rl_ateco, rl_ateco_macro, rl_ateco_settore, saldo, sede_op_ateco, sede_op_comune, sede_op_indirizzo, sede_op_provincia, somm, tipo_contratto, tipo_orario, cittadinanza, iso3, contientne, aggregazione, provincia, sigla_prov, comune_istat, codice_istat, SLL_codice, SLL_nome, contratto, erroriEta, errori_qualifica, qualifica_2_digit, erroriCF, errori. For the purpose of data exploration we will focus only on the following: CF, anno, eta, genere, iso3, professione, qualifica, saldo.

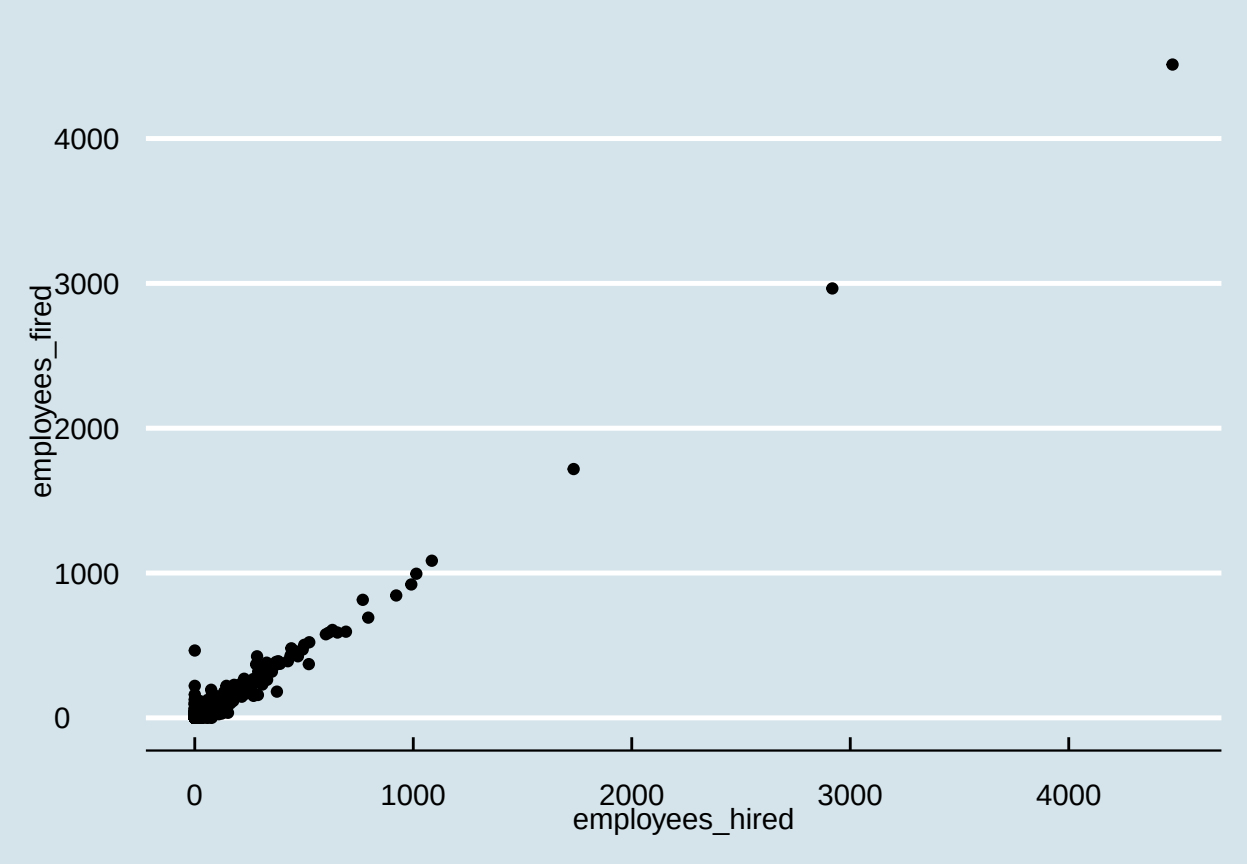
```
empl <- empl %>%
  select( one_of(some_features) ) %>%
  rename( year = anno)

empl_flows <- empl %>% select( c(CF, saldo, year)) %>%
  mutate(hf = factor(saldo))%>%
  mutate(hf=recode(hf,`-1`="fired",`1`="hired"))%>%
  group_by(CF,hf, year) %>%
  summarize(hiredfired=  sum(saldo) ) %>%
  pivot_wider( names_from = hf, values_from = hiredfired) %>%
  replace(is.na(.), 0) %>%
  mutate(turnover = hired-fired) %>%
  mutate(net = hired+fired)
```

```
## `summarise()` has grouped output by 'CF', 'hf'. You can override using the `.groups` argument.
```

```
employees_hired = log10(empl_flows$hired)
employees_fired=  log10(-empl_flows$fired)
ggplot(empl_flows, aes(x=employees_hired, y=employees_fired))+
    geom_point()
```




```
employees_hired = (empl_flows$hired)
employees_fired=  (-empl_flows$fired)
ggplot(empl_flows, aes(x=employees_hired, y=employees_fired))+
    geom_point()
```



> TODO import, calculate net saldo and turnover, divide companies in quartiles

TODO improve formatting tables with library(kableExtra) %>% kable()

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