Preprocessing

```
#echo "rmarkdown::render('preprosessing.Rmd', clean=TRUE,output_format='pdf_document',output_file='prep
#setwd(dir ="/home/ilpo/Paavo/src")
```

```
load("../data/paavodata.rdata")
load("../data/ilposdata.rdata")
```

```
ilposdata %>% mutate(zip = recode(zip,
                            #old value = new value
                            "106001" = "10600",
                            "201001" = "20100",
                            "216001"=  "21600",
                            "257001" = "25710",
                             "27001" = "27100",
                             "41301"= "41310" ,
                             "61001"="61100",
                             "651001"= "61100",
                             "669001"="66900",
                             "685001" = "68500",
                             "686001"= "68600",
                             "686201"= "68620",
                             "651001"= "06500",
                             "651001" = "65100",
                             "652001" = "65200",
                              "669001"= "66900",
                              "686001" = "68600",
                              "688001" = "68600",
                               "688001" = "68600"))
```

```
## # A tibble: 1,537,680 x 13
##    donor Site  dateonly   status donat_phleb    Hb gender aborh zip     age
##    <fct> <fct> <date>     <fct>  <fct>       <dbl> <fct>  <fct> <fct> <int>
##  1 DR00~ L3149 2018-10-08 -      K             141 Women  A Rh~ 90570    19
##  2 DR00~ L3149 2018-10-08 -      K             156 Men    O Rh~ 90530    21
##  3 DR00~ L3149 2018-10-08 -      K             156 Men    A Rh~ 90560    22
##  4 DR00~ L0564 2019-01-17 -      K             163 Men    A Rh~ 90560    22
##  5 DR00~ L3149 2018-10-08 R      K             127 Women  A Rh~ 90570    20
##  6 DR00~ L3149 2019-01-14 E      *             138 Women  A Rh~ 90570    20
##  7 DR00~ L3149 2018-10-08 E      *             144 Women  ""    90550    19
##  8 DR00~ L3149 2018-10-08 -      K             140 Women  O Rh~ 90530    21
##  9 DR00~ L3149 2018-10-08 -      K             128 Women  B Rh~ 90500    19
## 10 DR00~ L3149 2018-10-08 -      K             153 Women  AB R~ 90530    20
## # ... with 1,537,670 more rows, and 3 more variables: age.group <fct>,
## #   Hb_deferral <fct>, FirstEvent <lgl>
```

```
ilposdata$zip <- gsub('.*NA.*',NA,ilposdata$zip)
ilposdata <- ilposdata %>% filter(!is.na(zip))
```

Numbers of donors per zip and number of donations per zip

```
prepocessing <- ilposdata %>%
  mutate(Year = year(dateonly)) %>%
  filter(donat_phleb == "K") %>%
  filter(Year == 2017 | Year == 2018) %>%
    count(donor, Year,zip) %>%
    count(Year, zip) %>%
    rename(nb_donors_per_zip = nn)

ilposdata %>% filter(zip == '00120') %>%  filter(donat_phleb == "K") %>% mutate(year=year(dateonly)) %>%
```

```
## # A tibble: 118 x 2
## # Groups:   donor [118]
##    donor          n
##    <fct>      <int>
##  1 DR00935355     3
##  2 DR00949601     2
##  3 DR00905639     1
##  4 DR00669346     3
##  5 DR00635694     2
##  6 DR00438986     2
##  7 DR00472274     2
##  8 DR00756314     2
##  9 DR00184104     3
## 10 DR00330564     1
## # ... with 108 more rows
```

```
test <-ilposdata %>%
  mutate(Year = year(dateonly)) %>%
  filter(donat_phleb == "K") %>%
filter(Year == 2017| Year== 2018) %>%
count(Year,zip) %>%
rename(nb_donations_per_zip=n)

ilposdata %>% filter(zip == '00120') %>%  filter(donat_phleb == "K") %>% mutate(year=year(dateonly)) %>%
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   178
```

```
  preprosessing <-left_join(prepocessing,test,
  by = c("zip", "Year"))
```

nb_first_time_donors

```
firstevent <- ilposdata %>%
 mutate(Year = year(dateonly)) %>%
  filter(donat_phleb == "K") %>%
filter(Year == 2017| Year== 2018) %>%
select(zip,FirstEvent,Year) %>%
filter(FirstEvent == TRUE) %>%
group_by(zip,Year) %>%
summarise(nb_first_time_donors= n()) %>%
ungroup()

ilposdata %>% filter(zip == '00120') %>%  filter(donat_phleb == "K") %>% filter(FirstEvent== TRUE) %>% r
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    23
```

#nb_repeat_donors

```
# repeatedevent <- ilposdata %>%
#  mutate(Year = year(dateonly)) %>%
#  filter(Year == 2017| Year== 2018) %>%
#  filter(donat_phleb == "K") %>%
#   filter(FirstEvent == FALSE)
#
# repeatedevent <- repeatedevent %>%  group_by(donor, Year) %>%
#  filter(dateonly == min(dateonly))
#
#
# repeatedevent <- repeatedevent %>%  ungroup() %>%
# count(zip, Year) %>%
#   rename(nb_repeat_donors= n)

#A simpler way is to substrat first time donors from total donors

preprocessed <- left_join(preprosessing ,firstevent,
  by = c("zip", "Year")) %>% mutate( nb_repeat_donors= nb_donors_per_zip - nb_first_time_donors)

preprocessed$nb_first_time_donors[is.na(preprocessed$nb_first_time_donors)] <- 0
preprocessed$nb_repeat_donors[is.na(preprocessed$nb_repeat_donors)] <- 0
```

```
# events <- left_join(firstevent,repeatedevent,
# by = c("zip","Year"))
```

```
# preprocessed <- left_join(preprosessing ,events,
#   by = c("zip", "Year"))
```

#joining the data with Paavodata

```r
preprosessed_paavo <- paavodata %>%
  rename(zip = pono, Year= vuosi) %>%
  filter(Year == 2019) %>%
  mutate(eligible_population = he_18_19 + he_20_24 + he_25_29 + he_30_34 + he_40_44 +
                                he_45_49 + he_50_54 + he_55_59 + he_60_64 + he_65_69) %>%
  dplyr::select(-Year)

 preprosessed_paavo <-
  right_join(preprosessed_paavo, preprocessed, by = c("zip")) %>%
  dplyr::select(zip, Year, eligible_population, nb_donors_per_zip,nb_donations_per_zip, nb_first_time_do
pt_tyott, pt_tyoll,  ko_ika18y, hr_tuy) %>%
    rename (unemployed = pt_tyott,
            employed = pt_tyoll,
            medianincome= hr_mtu,
            averageincome= hr_ktu,
            population18= ko_ika18y,
            bachelor_degree= ko_al_kork,
            masters_degree= ko_yl_kork,
           averageincome= hr_ktu,
           medianincome= hr_mtu) %>%
mutate( prop_donors= nb_donors_per_zip/eligible_population,
        nb_donation_per_act_donor= nb_donations_per_zip/nb_donors_per_zip,
        prop_new_donors= nb_first_time_donors/eligible_population,
        prop_repeat_donors= nb_repeat_donors/eligible_population,
         higher_education =bachelor_degree+masters_degree,
  proportion_inhabitants_with_higher_education= higher_education/eligible_population)


#Drop post-codes with no data

preprosessed_paavo <- preprosessed_paavo %>% filter(!is.na(eligible_population))


#Are there NA's left
preprosessed_paavo[apply(preprosessed_paavo,1,function(x){any(is.na(x))}),]
```

```
##        zip Year eligible_population nb_donors_per_zip nb_donations_per_zip
## 667  22930 2017                  21                2                    3
## 1515 57120 2017                  23                3                    9
## 1900 69970 2017                  21                1                    1
## 2040 74980 2017                  25                1                    1
## 2365 90590 2017                 178                2                    2
## 3276 22930 2018                  21                2                    4
## 4123 57120 2018                  23                3                   10
## 4510 69970 2018                  21                1                    1
##      nb_first_time_donors nb_repeat_donors medianincome averageincome
## 667                     0                0        26929         26583
## 1515                    0                0           NA            NA
## 1900                    0                0           NA            NA
## 2040                    0                0           NA            NA
## 2365                    1                1           NA            NA
## 3276                    0                0        26929         26583
## 4123                    0                0           NA            NA
## 4510                    0                0           NA            NA
```

```
##                nimi bachelor_degree masters_degree unemployed employed
## 667           Fiskö              NA             NA          0       10
## 1515     Savonlinna              NA             NA         NA       NA
## 1900    Salamajärvi              NA             NA          3        7
## 2040  Tihilänkangas               0              0          5       15
## 2365 Teknologiakylä              18              3         NA       NA
## 3276          Fiskö              NA             NA          0       10
## 4123     Savonlinna              NA             NA         NA       NA
## 4510    Salamajärvi              NA             NA          3        7
##      population18 hr_tuy prop_donors nb_donation_per_act_donor
## 667            29     31  0.09523810                 1.500000
## 1515           26     24  0.13043478                 3.000000
## 1900           28     28  0.04761905                 1.000000
## 2040           30     29  0.04000000                 1.000000
## 2365          183      1  0.01123596                 1.000000
## 3276           29     31  0.09523810                 2.000000
## 4123           26     24  0.13043478                 3.333333
## 4510           28     28  0.04761905                 1.000000
##      prop_new_donors prop_repeat_donors higher_education
## 667      0.000000000        0.000000000               NA
## 1515     0.000000000        0.000000000               NA
## 1900     0.000000000        0.000000000               NA
## 2040     0.000000000        0.000000000                0
## 2365     0.005617978        0.005617978               21
## 3276     0.000000000        0.000000000               NA
## 4123     0.000000000        0.000000000               NA
## 4510     0.000000000        0.000000000               NA
##      proportion_inhabitants_with_higher_education
## 667                                            NA
## 1515                                           NA
## 1900                                           NA
## 2040                                    0.0000000
## 2365                                    0.1179775
## 3276                                           NA
## 4123                                           NA
## 4510                                           NA
```

```
summary(preprosessed_paavo)
```

```
##      zip                 Year      eligible_population nb_donors_per_zip
##  Length:5209        Min.   :2017   Min.   :   20       Min.   :   1.00
##  Class :character   1st Qu.:2017   1st Qu.:  147       1st Qu.:   3.00
##  Mode  :character   Median :2017   Median :  380       Median :  11.00
##                     Mean   :2017   Mean   : 1233       Mean   :  43.34
##                     3rd Qu.:2018   3rd Qu.: 1473       3rd Qu.:  46.00
##                     Max.   :2018   Max.   :18536       Max.   :1111.00
##
##  nb_donations_per_zip nb_first_time_donors nb_repeat_donors
##  Min.   :   1.00      Min.   :  0.000      Min.   :  0.00
##  1st Qu.:   6.00      1st Qu.:  0.000      1st Qu.:  0.00
##  Median :  19.00      Median :  1.000      Median :  6.00
##  Mean   :  74.26      Mean   :  5.617      Mean   : 35.83
##  3rd Qu.:  78.00      3rd Qu.:  5.000      3rd Qu.: 40.00
##  Max.   :1898.00      Max.   :193.000      Max.   :932.00
##
```

```
##   medianincome   averageincome        nimi           bachelor_degree
##  Min.   :10469   Min.   : 12124   Length:5209        Min.   :   0
##  1st Qu.:18060   1st Qu.: 20426   Class :character   1st Qu.:  14
##  Median :19947   Median : 22065   Mode  :character   Median :  46
##  Mean   :20286   Mean   : 22788                      Mean   : 197
##  3rd Qu.:22210   3rd Qu.: 24196                      3rd Qu.: 213
##  Max.   :35612   Max.   :100488                      Max.   :4036
##  NA's   :6       NA's   :6                           NA's   :6
##  masters_degree     unemployed        employed        population18
##  Min.   :   0.0   Min.   :   0.0   Min.   :    5.0   Min.   :   26
##  1st Qu.:   8.0   1st Qu.:  14.0   1st Qu.:   97.0   1st Qu.:  204
##  Median :  27.0   Median :  37.0   Median :  266.5   Median :  516
##  Mean   : 181.2   Mean   : 132.4   Mean   :  862.1   Mean   : 1670
##  3rd Qu.: 148.0   3rd Qu.: 149.0   3rd Qu.: 1023.8   3rd Qu.: 2013
##  Max.   :5704.0   Max.   :2654.0   Max.   :13006.0   Max.   :24408
##  NA's   :6        NA's   :3        NA's   :3
##      hr_tuy       prop_donors        nb_donation_per_act_donor
##  Min.   :    1   Min.   :0.0006667   Min.   :1.000
##  1st Qu.:  207   1st Qu.:0.0198610   1st Qu.:1.500
##  Median :  519   Median :0.0303616   Median :1.704
##  Mean   : 1664   Mean   :0.0317566   Mean   :1.734
##  3rd Qu.: 2021   3rd Qu.:0.0407252   3rd Qu.:1.929
##  Max.   :24161   Max.   :0.1333333   Max.   :6.000
##
##  prop_new_donors    prop_repeat_donors higher_education
##  Min.   :0.000000   Min.   :0.00000    Min.   :   0.0
##  1st Qu.:0.000000   1st Qu.:0.00000    1st Qu.:  22.0
##  Median :0.002469   Median :0.01786    Median :  76.0
##  Mean   :0.003258   Mean   :0.01810    Mean   : 378.2
##  3rd Qu.:0.005031   3rd Qu.:0.03180    3rd Qu.: 365.0
##  Max.   :0.071429   Max.   :0.12000    Max.   :8603.0
##                                        NA's   :6
##  proportion_inhabitants_with_higher_education
##  Min.   :0.0000
##  1st Qu.:0.1405
##  Median :0.1884
##  Mean   :0.2176
##  3rd Qu.:0.2708
##  Max.   :0.7696
##  NA's   :6
```

```
save(preprosessed_paavo,  file = paste0("preprocessed.",Sys.Date(),".RData"))
```