

Skill intensity over years

Contents

1	Wage Data	4
1.1	Missing variables	5
1.2	Update features	7
1.3	Distributions	7
1.4	Correlation	8
1.5	Can I link plants to firms?	9

```
library(VNFirmSurvey)
library(data.table)
library(ggplot2)
library(DataExplorer)
library(here)
#> here() starts at /Users/nghiem/Documents/data-projects/VNFirmSurvey
library(haven)
```

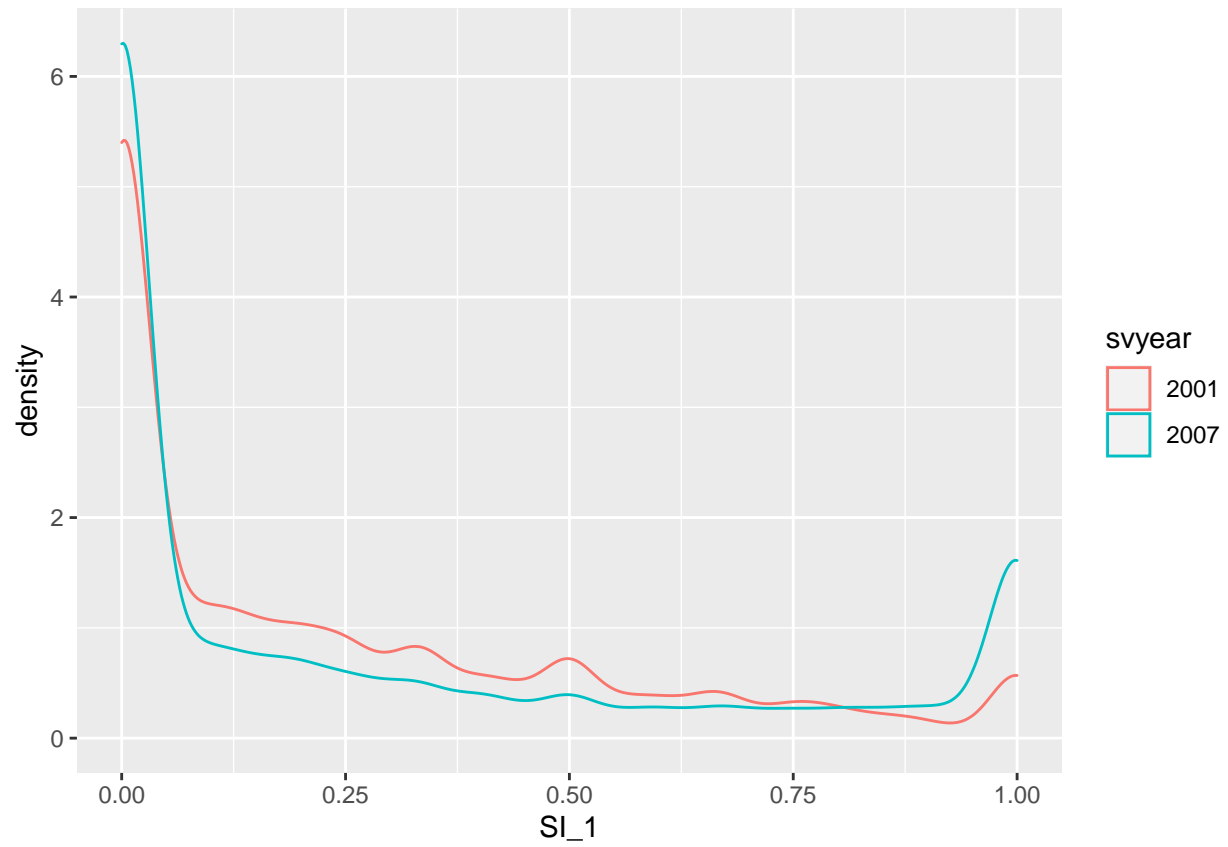
```
data("skill_07_dta")
data("skill_01_dta")

dta_list <- list(skill_07_dta, skill_01_dta)

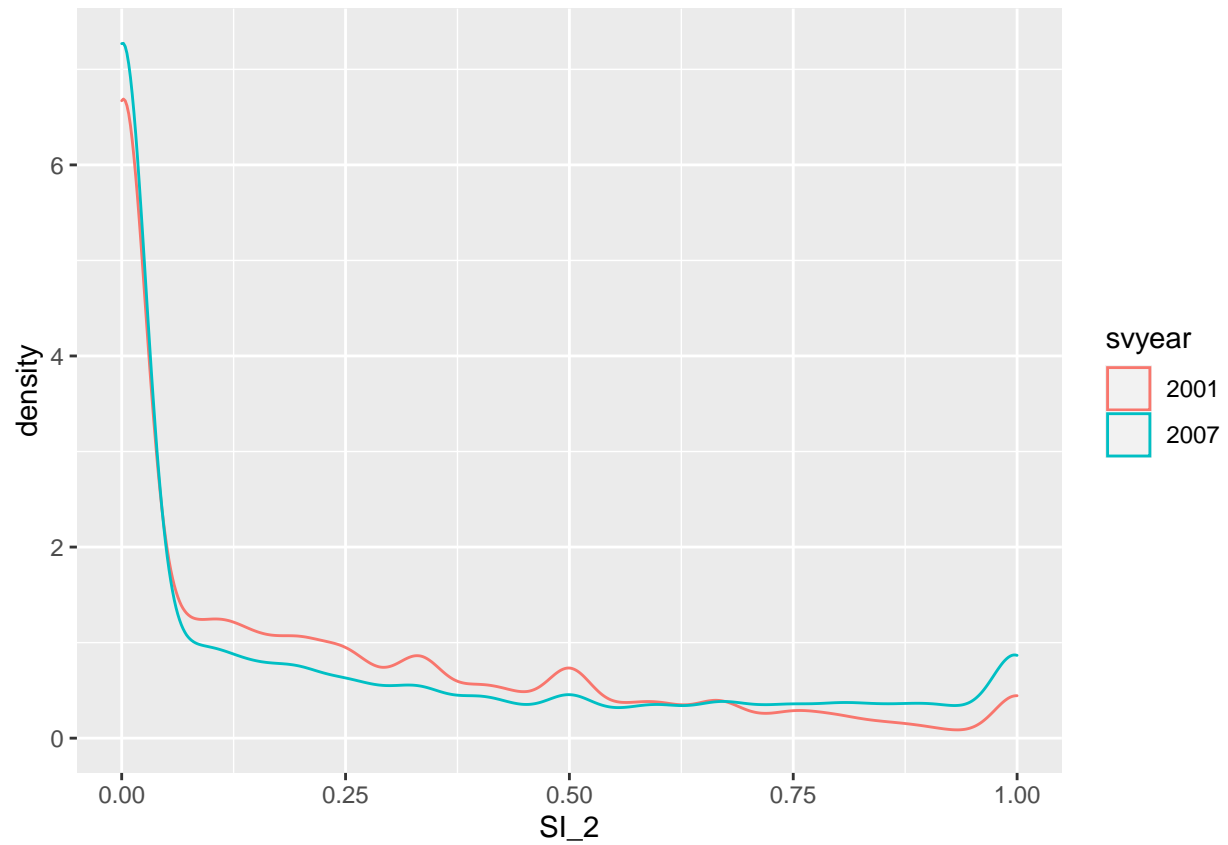
stacked_data <- rbindlist(dta_list)
```

```
stacked_data$svyear <- as.factor(stacked_data$svyear)

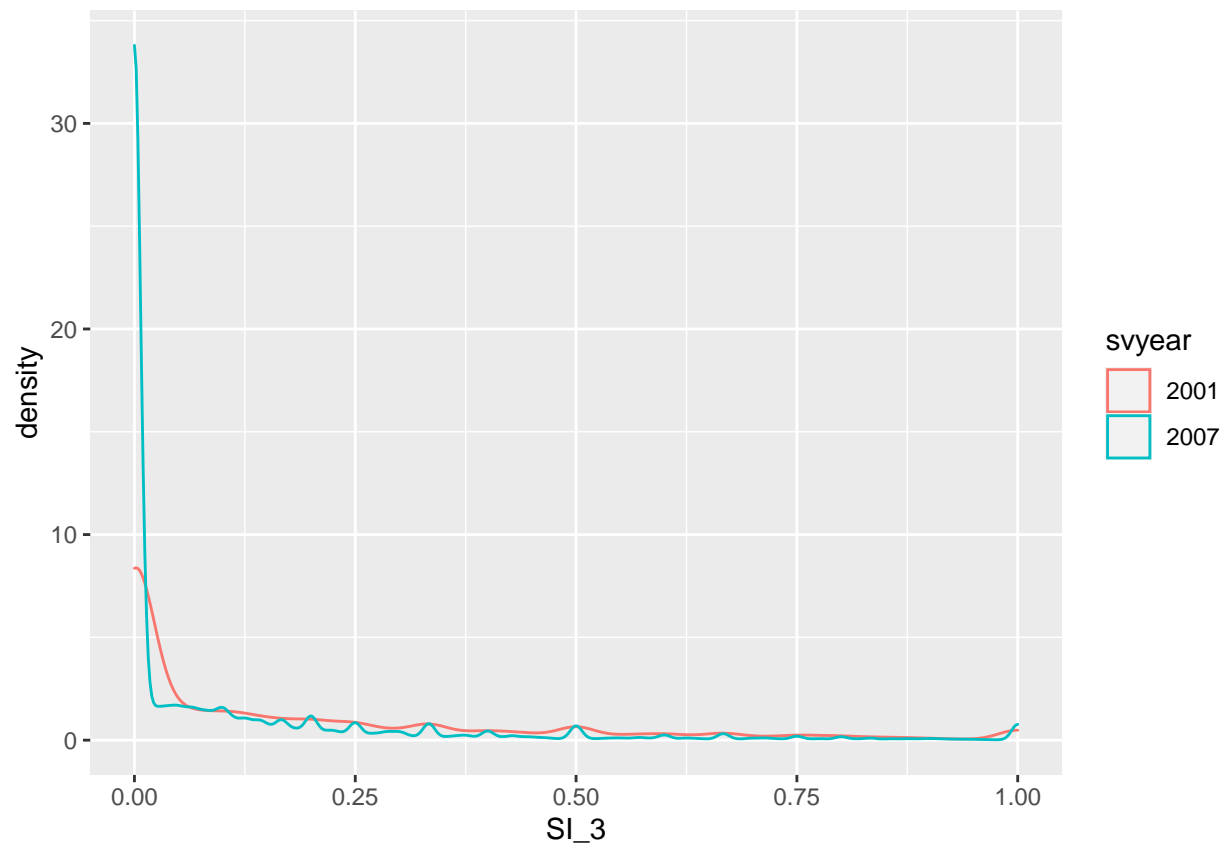
ggplot(stacked_data, aes(x = SI_1, col = svyear)) +
  geom_density()
```



```
ggplot(stacked_data, aes(x = SI_2, col = svyear)) +  
  geom_density()
```



```
ggplot(stacked_data, aes(x = SI_3, col = svyear)) +  
  geom_density()
```



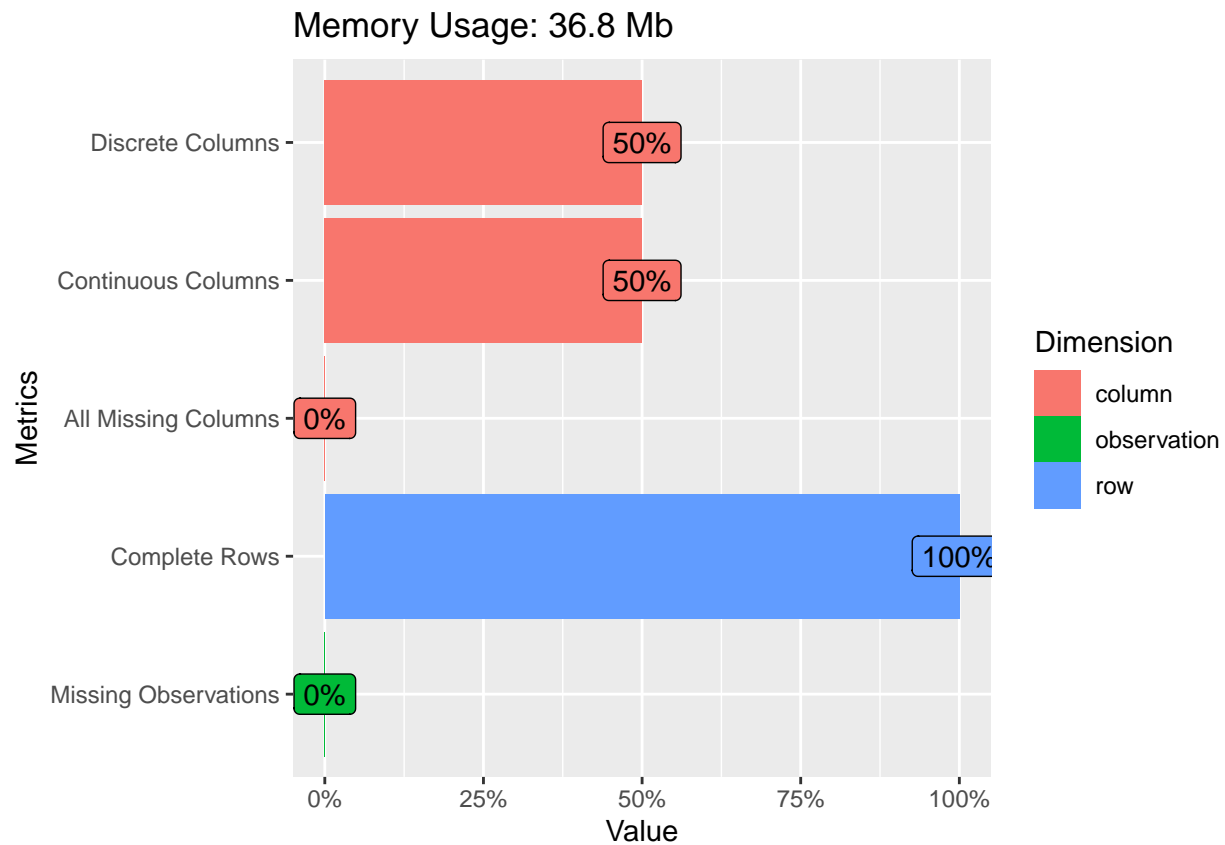
1 Wage Data

```
dta_list <- list(here("inst", "extdata",  
                     "Stata_2001",  
                     "dn2001.dta"),  
                here("inst", "extdata",  
                     "Stata_2007",  
                     "dn2007.dta") )
```

```
wage_dta <- getWage(dta_list)
```

Let's take a first look at the data

```
plot_intro(wage_dta)
```

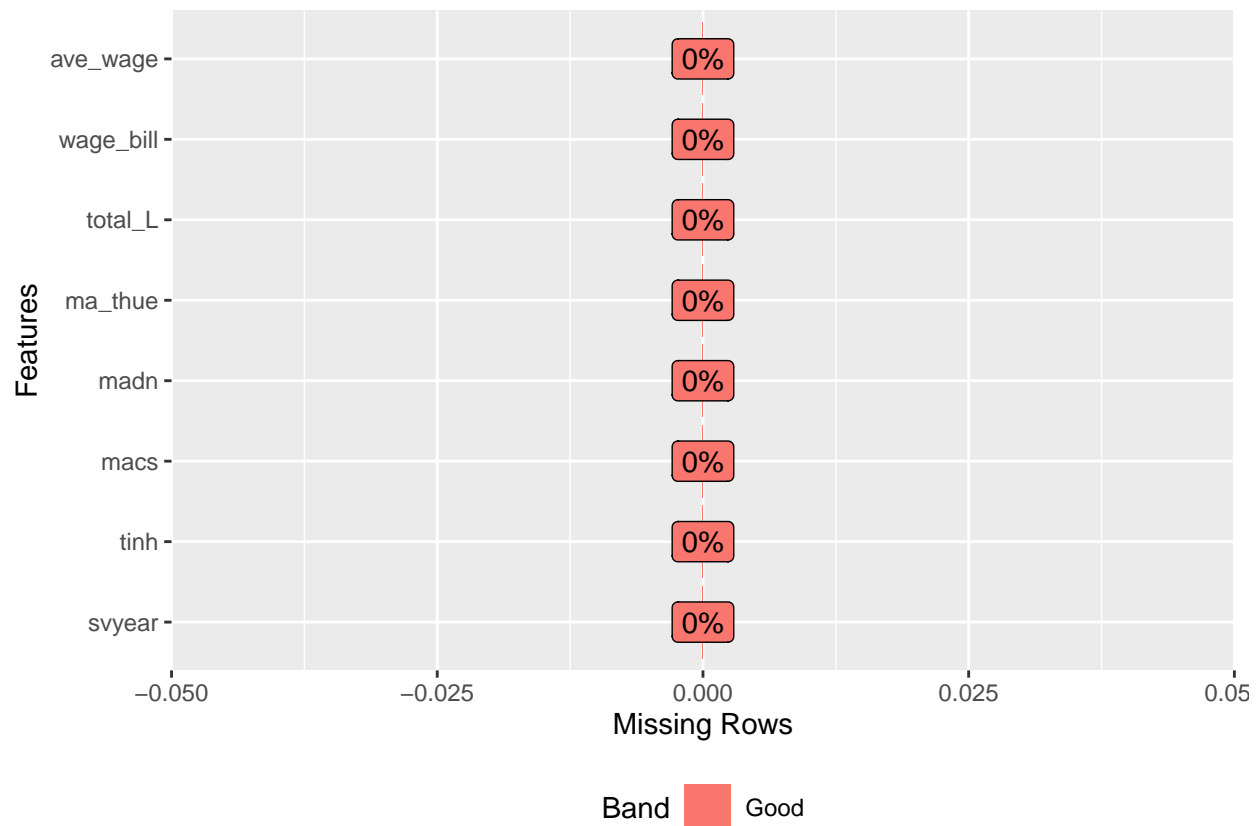


1.1 Missing variables

If I use ld13 for 2001, I have more than 3000 missing observations for total labor while only 8 for ldc11.

Furthermore, 9.6% of the firms in 2001 miss total compensation data, while only 0.12% miss data

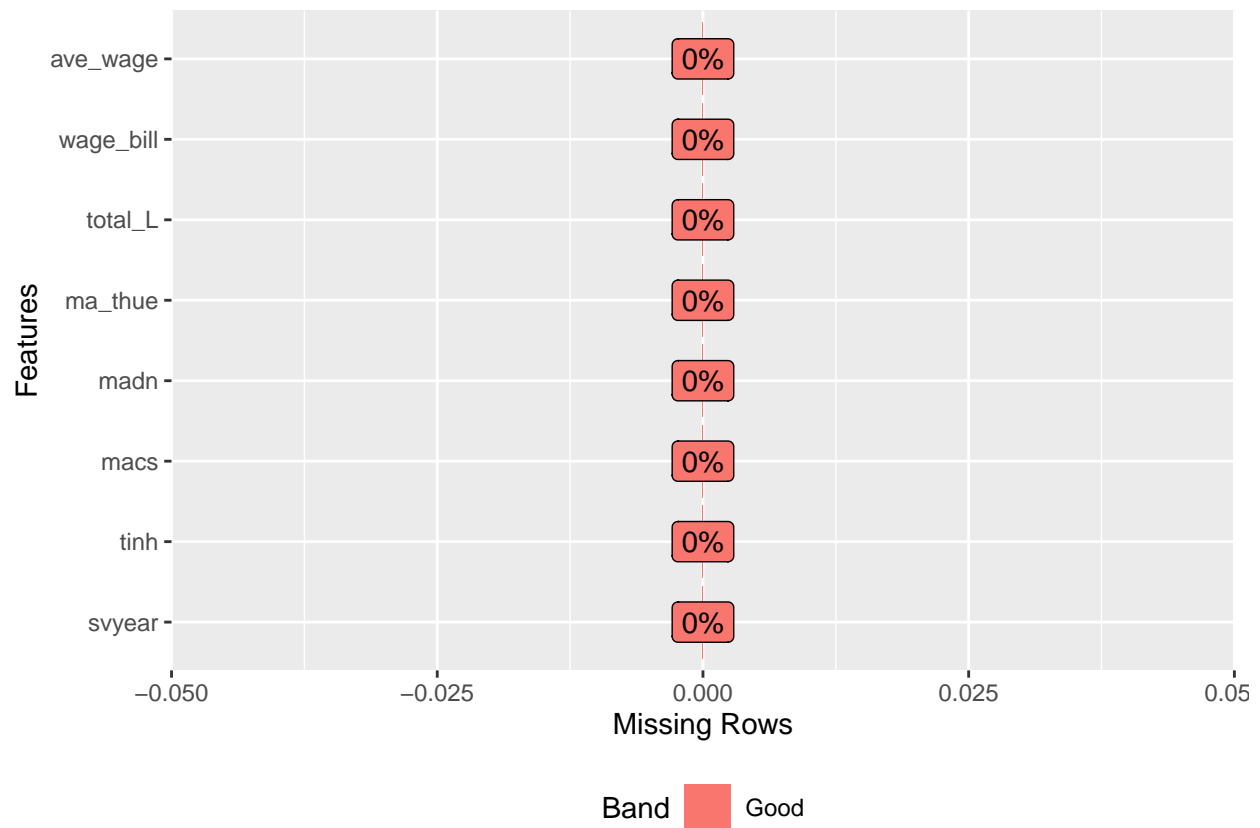
```
plot_missing(wage_dta)
```



```
profile_missing(wage_dta)
#>   feature num_missing pct_missing
#> 1:  svyear           0           0
#> 2:   tinh           0           0
#> 3:   macs           0           0
#> 4:   madn           0           0
#> 5:  ma_thue         0           0
#> 6: total_L           0           0
#> 7: wage_bill        0           0
#> 8: ave_wage         0           0
```

```
wage_dta_na_free <- na.omit(wage_dta)

wage_dta_na_free <- wage_dta_na_free[total_L > 0]
plot_missing(wage_dta_na_free)
```

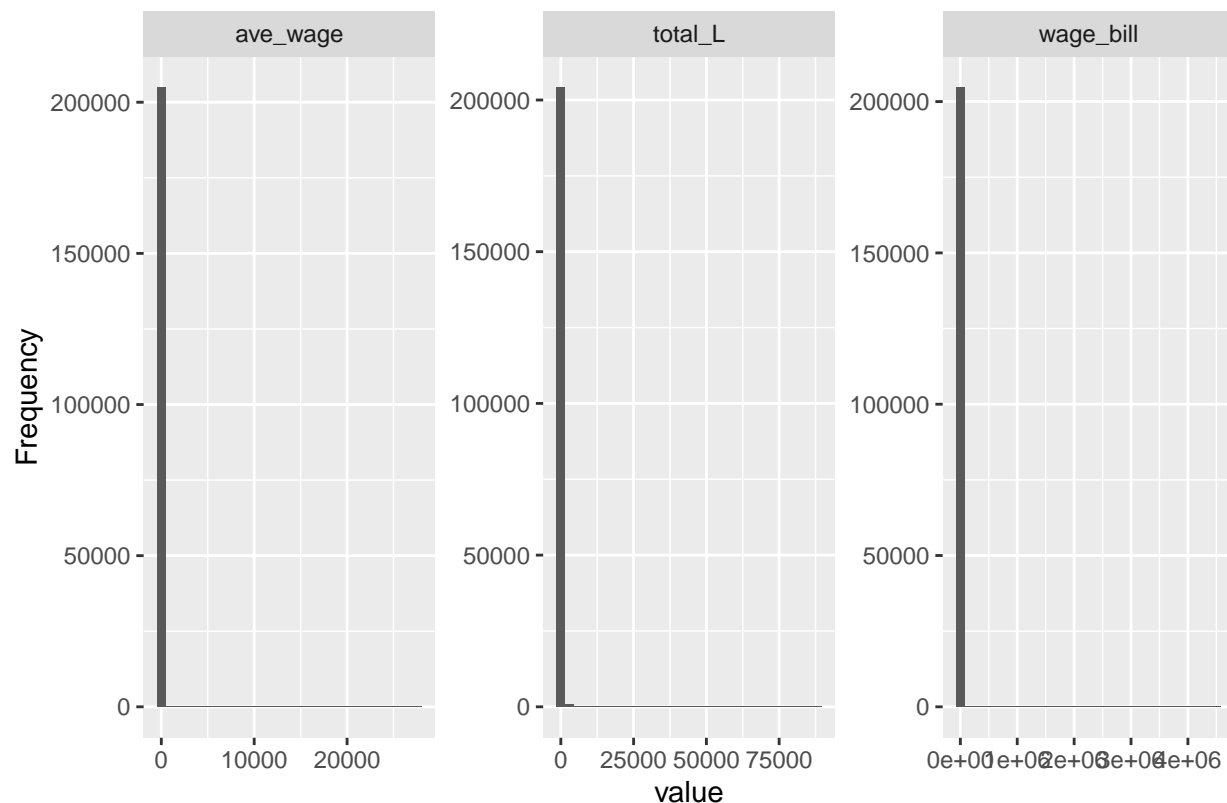


1.2 Update features

```
update_columns(wage_dta, c("tinh", "madn", "macs", "ma_thue"), as.factor)
```

1.3 Distributions

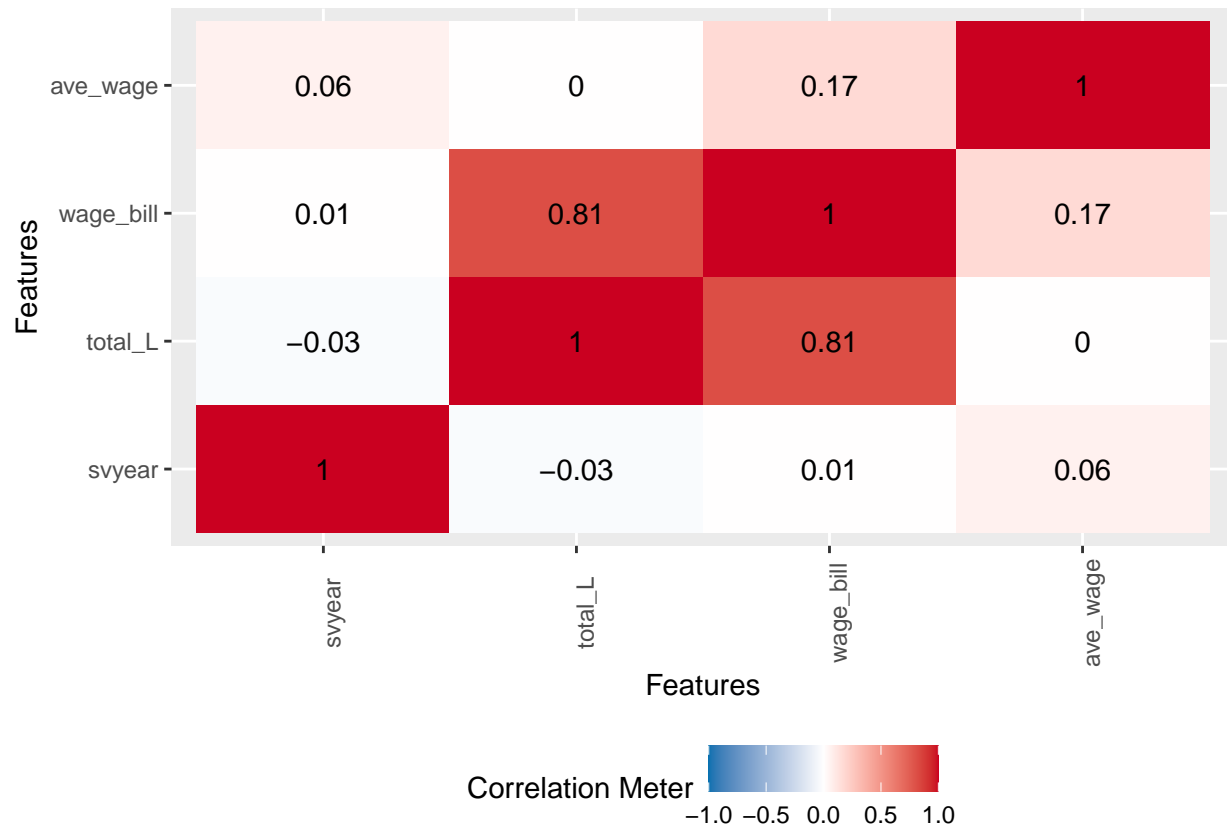
```
plot_histogram(wage_dta_na_free)
```



```
summary(wage_dta_na_free)
#>      suyear      tinh      macs      madn
#> Min.   :2001    79      : 45069    1      : 153    285      :    7
#> 1st Qu.:2007    01      : 24668    3      : 133   15827     :    5
#> Median :2007   701      : 10781    2      : 131    326      :    4
#> Mean   :2006   101      :  5872    4      : 129    421      :    4
#> 3rd Qu.:2007    31      :  4495    6      : 126    854      :    4
#> Max.   :2007    74      :  4382    5      : 122   5891      :    4
#>      (Other):109710 (Other):204183 (Other):204949
#>      ma_thue      total_L      wage_bill      ave_wage
#> Cha cã : 5761 Min.   : 1.00 Min.   : 0 Min.   : 0.000
#>      : 185 1st Qu.: 5.00 1st Qu.: 58 1st Qu.: 8.462
#> 260035750: 11 Median : 9.00 Median : 128 Median : 14.800
#> 010010068: 7 Mean   : 55.21 Mean   : 1210 Mean   : 18.873
#> 010036457: 6 3rd Qu.: 22.00 3rd Qu.: 349 3rd Qu.: 22.702
#> 010010442: 5 Max.   :88071.00 Max.   :4506542 Max.   :27600.000
#> (Other) :199002
```

1.4 Correlation

```
plot_correlation(wage_dta_na_free, maxcat = 5L)
#> Warning in dummify(data, maxcat = maxcat): Ignored all discrete features since
#> `maxcat` set to 5 categories!
```

1.5 Can I link plants to firms?

```
library(haven)
cn2007 <- (read_dta("/Volumes/GoogleDrive/My Drive/econ_datasets/Vietnam_VES/Data/Stata_2007_2009/Stata_2007_2009.dta"))
festive_exp <- (read_dta("/Volumes/GoogleDrive/My Drive/econ_datasets/Vietnam_VES/Data/Stata_2007_2009/Stata_2007_2009.dta",
                        encoding="latin1"))
names(festive_exp)

setDT(festive_exp)

festive_exp[!duplicated(macs), .(tinh, macs) ,by = .(madn)][order(madn)]

View(head(festive_exp))
```

It looks like this is a firm survey, and not a plant survey because the plants are listed under tencn, while madn and macs are firm identifiers.

find out across years what information I have at the plants level.