Study of French labour market and inequalities

L. Insolia, J. Kim and G. Yeghikyan

SNS

— Midterm results —

March 14, 2018

Objectives

- Structure of French labour market
- Inequalities (in terms of salary):
  - ages
  - gender
  - job categories
  - spatial distribution
- Firms' distribution
- Exploratory analyses

Methodology

INSEE data

- Population: age, sex and cohabitation mode
- Salary: job categories, age and sex (mean net salary per hour in €)
- Firms: number of firms for each size
- Geography: GPS location

for different geographical levels (communes, departments, towns) in 2014

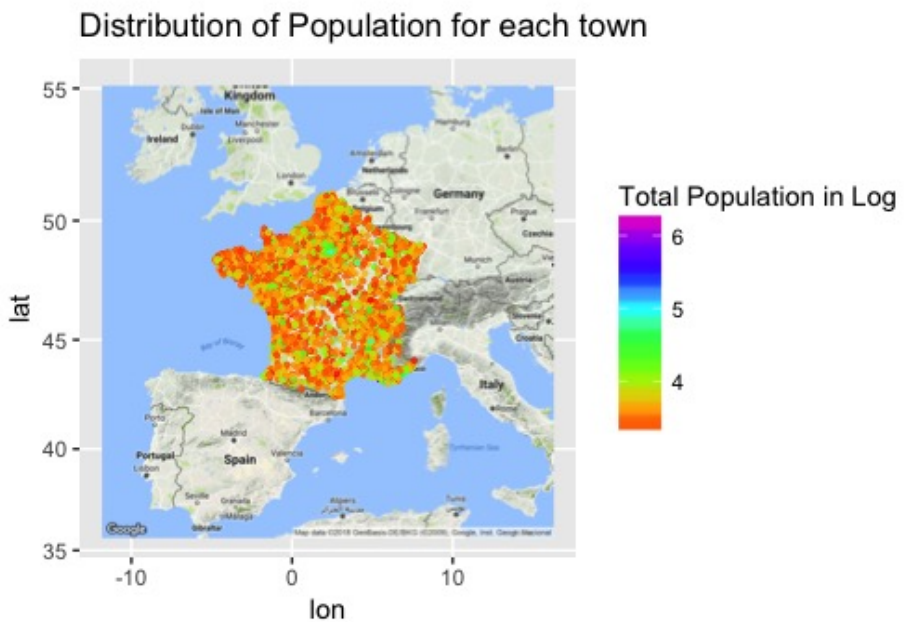
GitHub repo: https://github.com/LucaIns/TSL

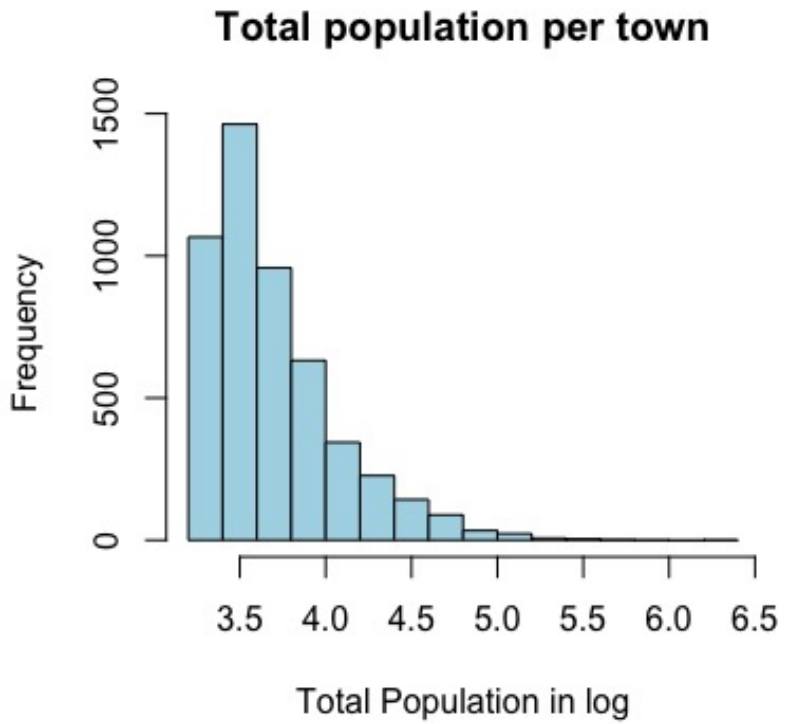
What has been done so far . . .

Pre-processing phase

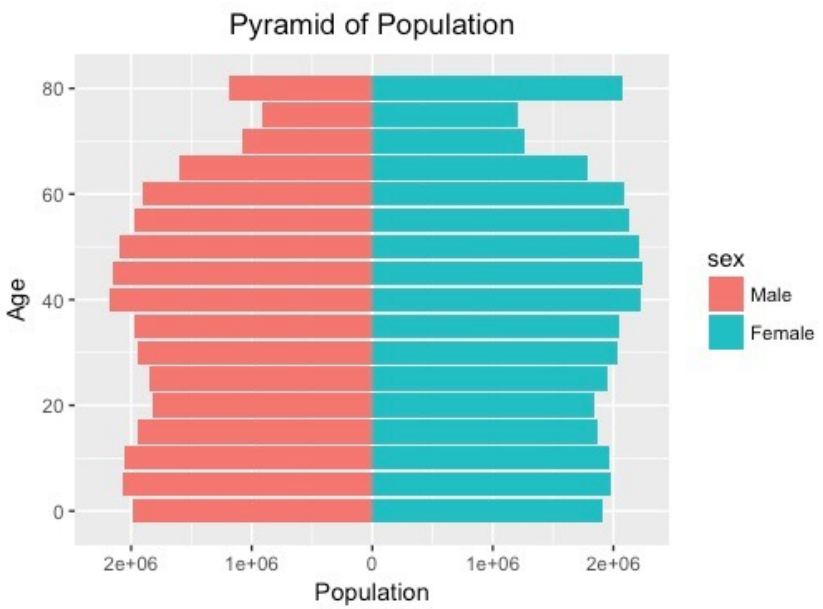
- Population: restructured the dataset and created new features
- Firms: categorized firms' sizes into 4 categories
- Geography: retrieved the missing data using Google API

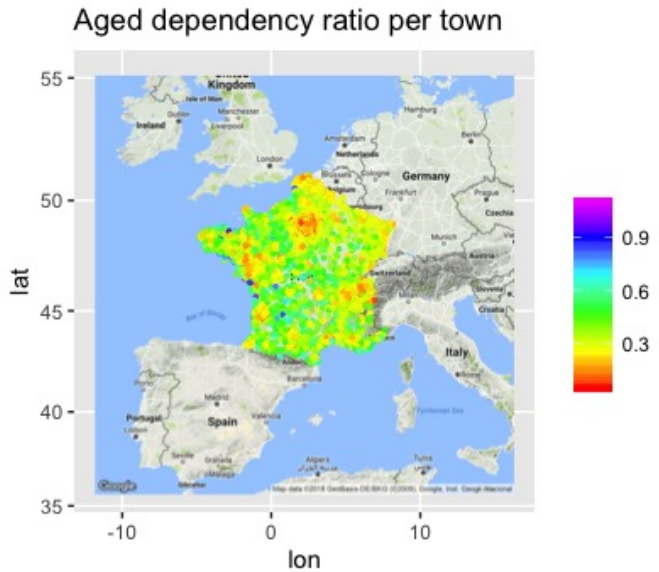
Distribution of population per town



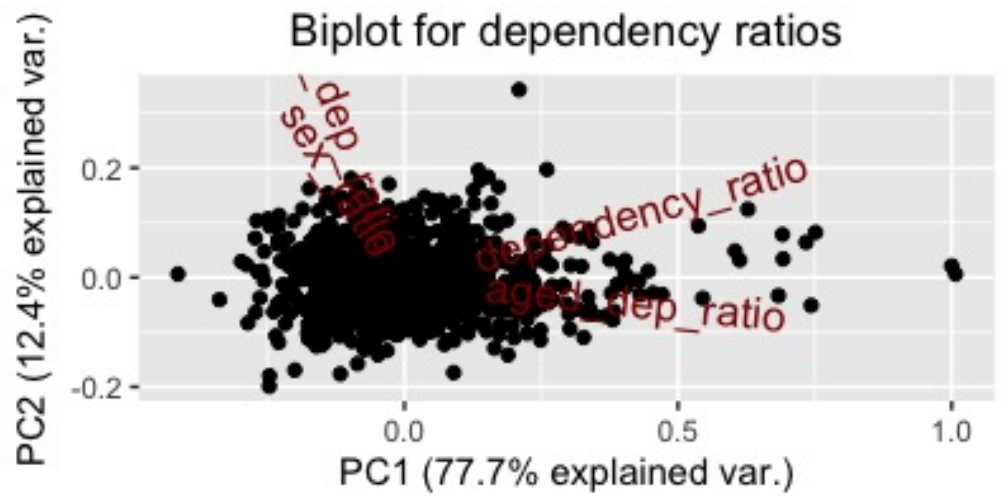


Population demographics

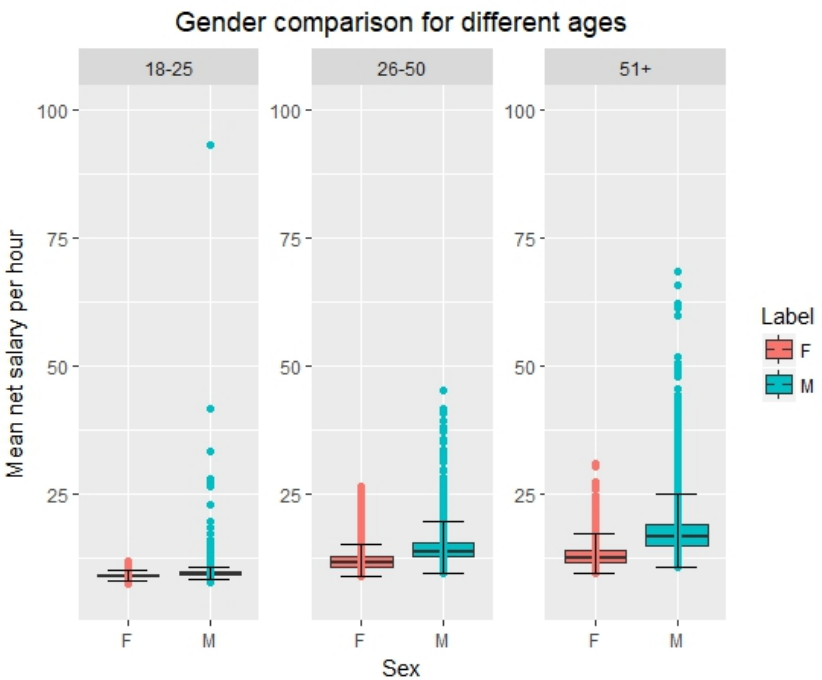


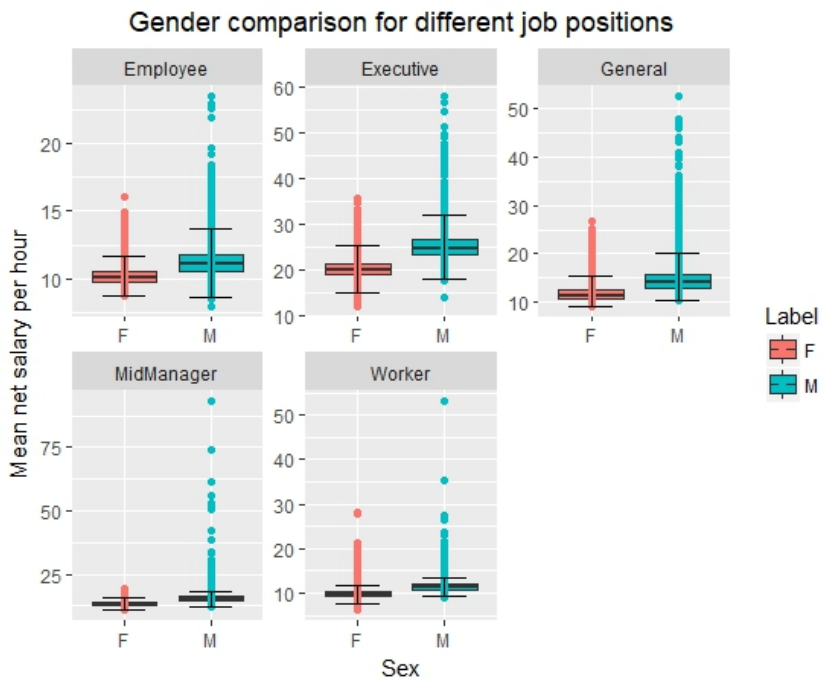


PCA

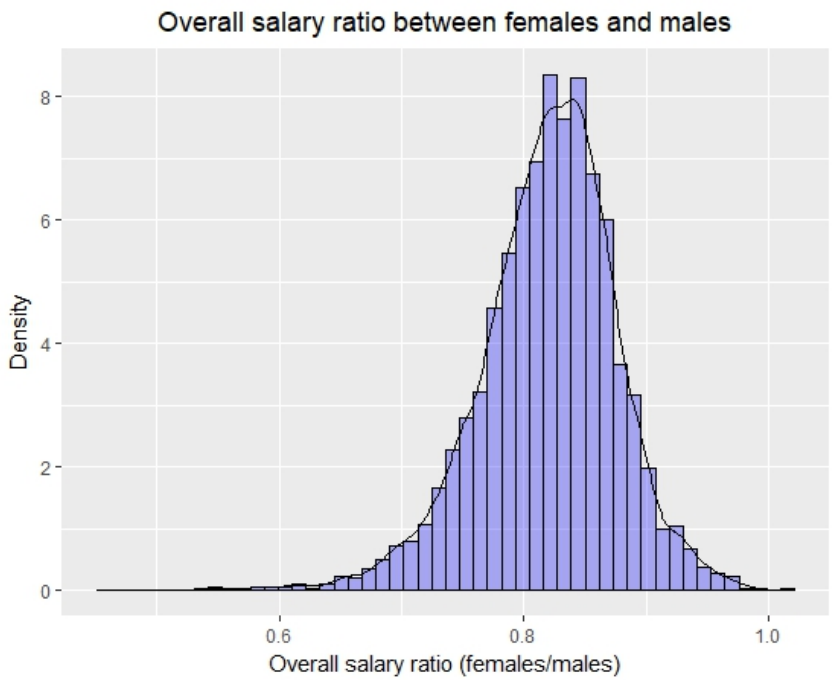

Biplot for dependency ratios
PC2 (12.4% explained var.)
PC1 (77.7% explained var.)
0.2
0.0
-0.2
0.0
0.5
1.0
dependency_ratio
age_dep_ratio

Inequality of salary





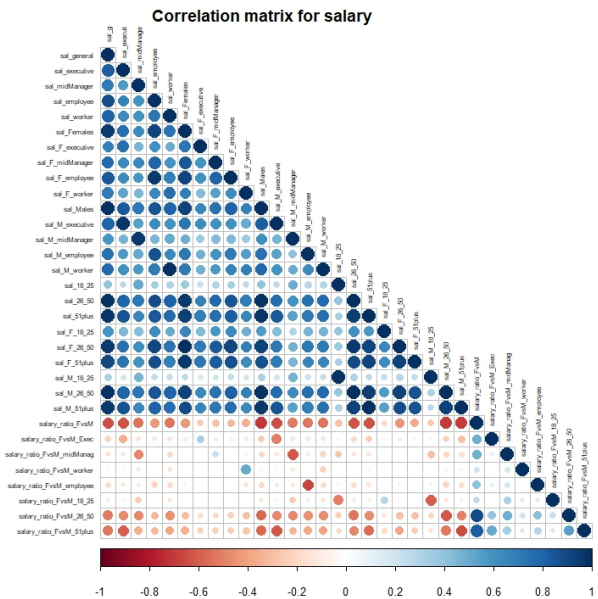
Inequality of salary

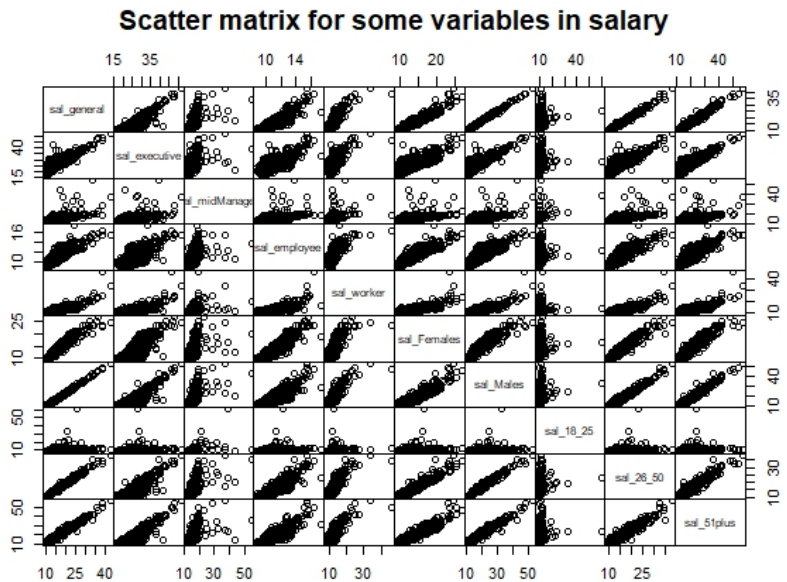


Inequality of salary

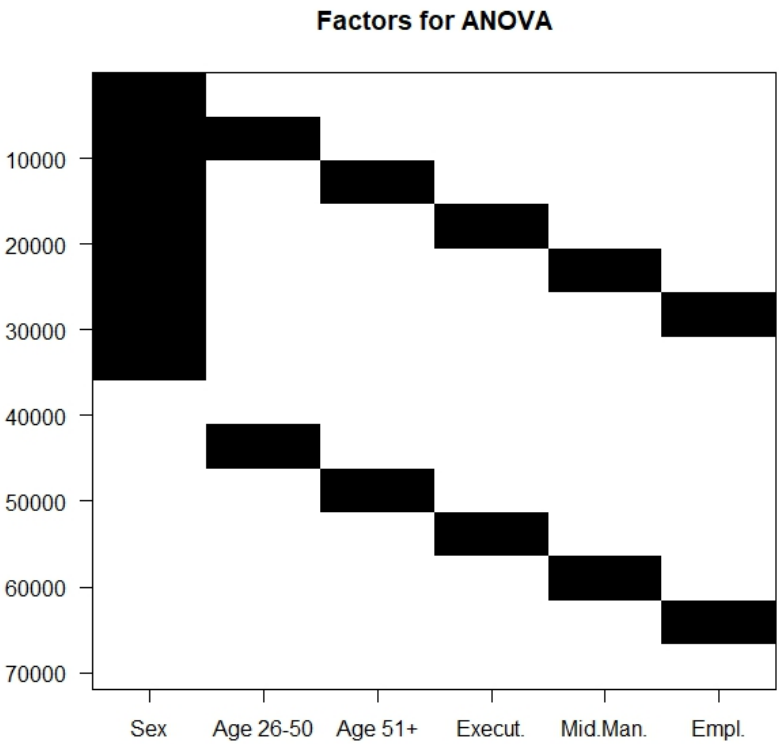
Inequality of salary

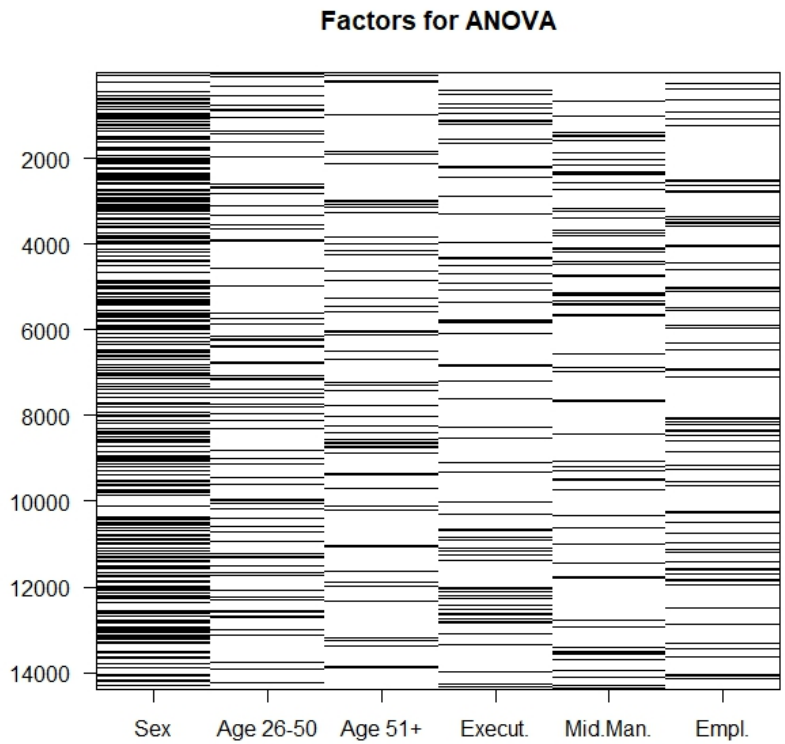
Bivariate relations



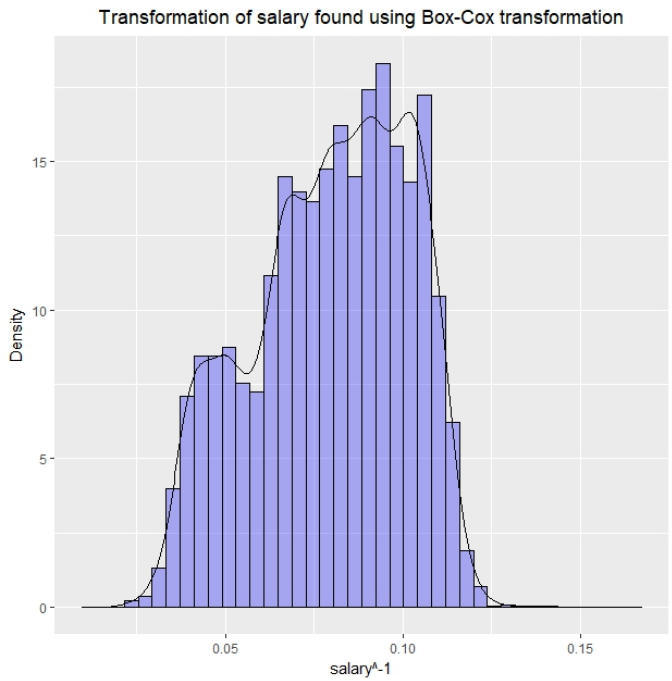


ANOVA


Factors for ANOVA
10000
20000
30000
40000
50000
60000
70000
Sex
Age 26-50
Age 51+
Execut.
Mid.Man.
Empl.

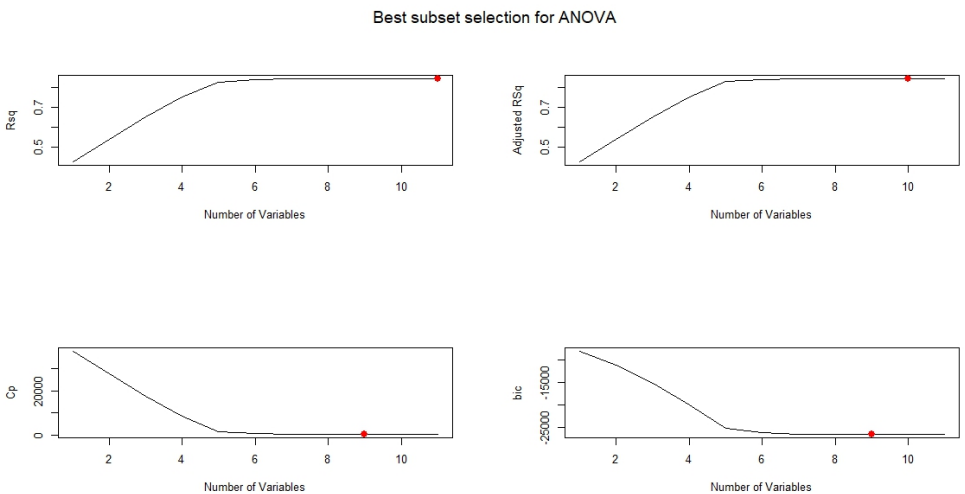

Factors for ANOVA
2000
4000
6000
8000
10000
12000
14000
Sex
Age 26-50
Age 51+
Execut.
Mid.Man.
Empl.

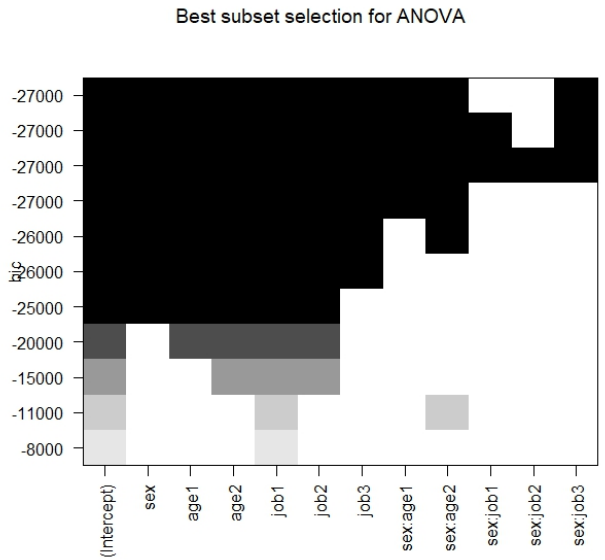
ANOVA



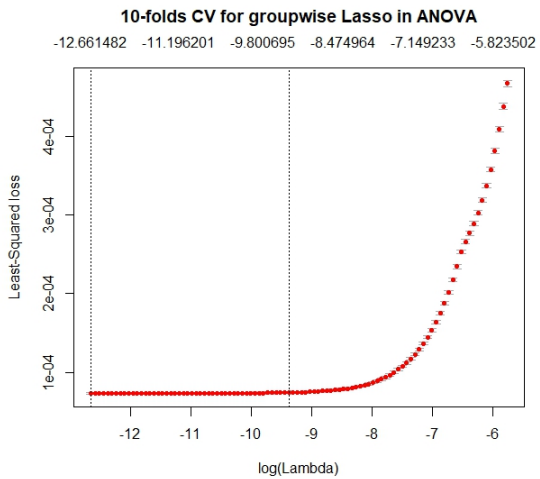

```
Call:
lm(formula = sal_y ~ sex + age + job + sex:age + sex:job)

Residuals:
      Min        1Q    Median        3Q       Max
-0.070217 -0.004526  0.000711  0.005522  0.057875

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.1060590  0.0001889  561.599  < 2e-16 ***
sex         -0.0106241  0.0002657  -39.989  < 2e-16 ***
age1        -0.0216544  0.0003289  -65.833  < 2e-16 ***
age2        -0.0278782  0.0003254  -85.667  < 2e-16 ***
job1        -0.0560085  0.0003239 -172.923  < 2e-16 ***
job2        -0.0303178  0.0003275  -92.566  < 2e-16 ***
job3        -0.0085416  0.0003254  -26.247  < 2e-16 ***
sex:age1    -0.0033048  0.0004648   -7.110 1.22e-12 ***
sex:age2    -0.0084595  0.0004613  -18.339  < 2e-16 ***
sex:job1     0.0005974  0.0004607    1.297    0.195
sex:job2     0.0002832  0.0004626    0.612    0.540
sex:job3     0.0026619  0.0004639    5.738 9.75e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008563 on 14369 degrees of freedom
Multiple R-squared:  0.8432,    Adjusted R-squared:  0.8431
F-statistic:  7026 on 11 and 14369 DF,  p-value: < 2.2e-16
```

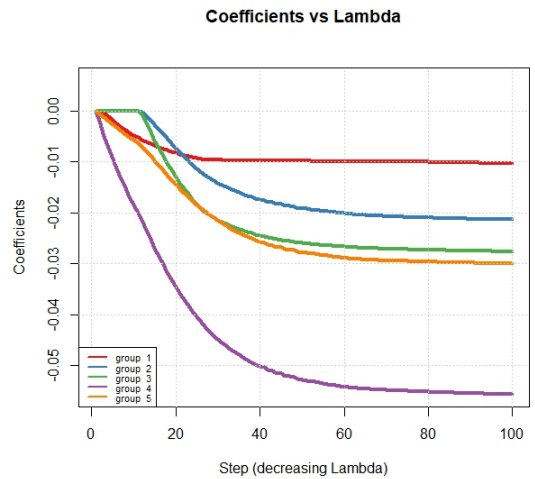
ANOVA BSS

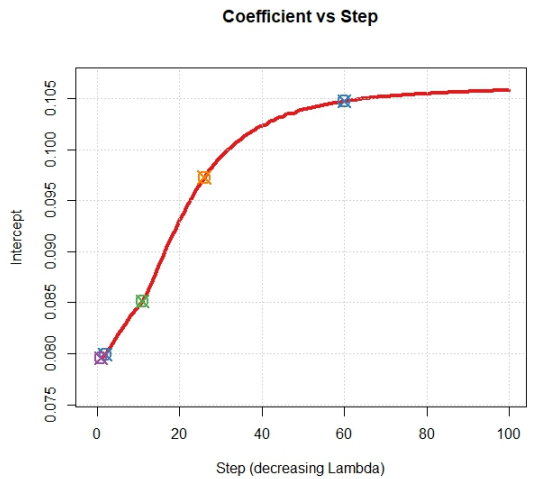

Best subset selection for ANOVA
Rsq
Adjusted RSq
Cp
bic
Number of Variables


Best subset selection for ANOVA
bic
(Intercept)
sex
age1
age2
job1
job2
job3
sex:age1
sex:age2
sex:job1
sex:job2
sex:job3

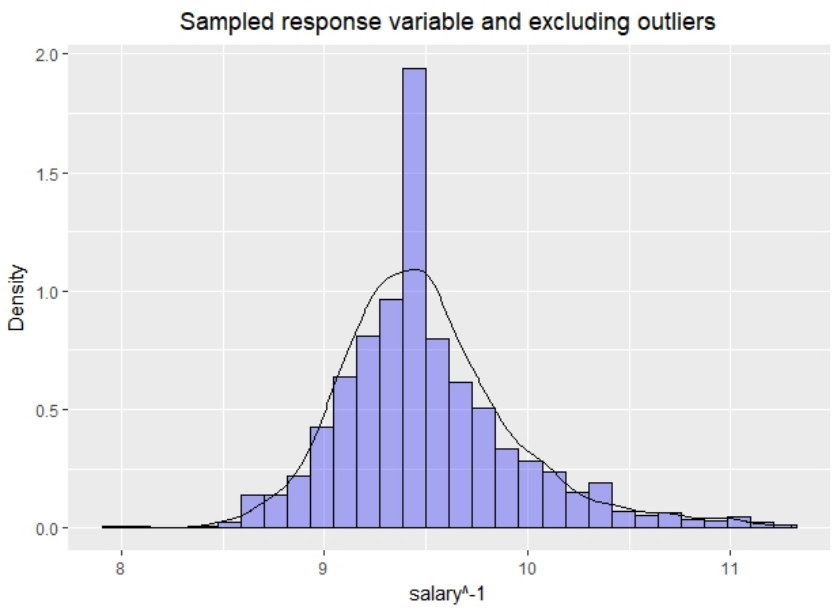
ANOVA GW-Lasso

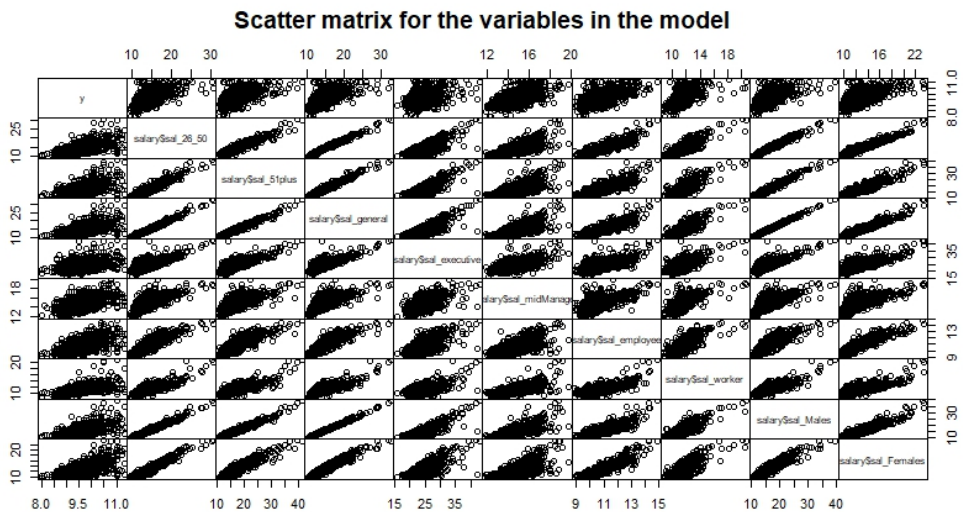




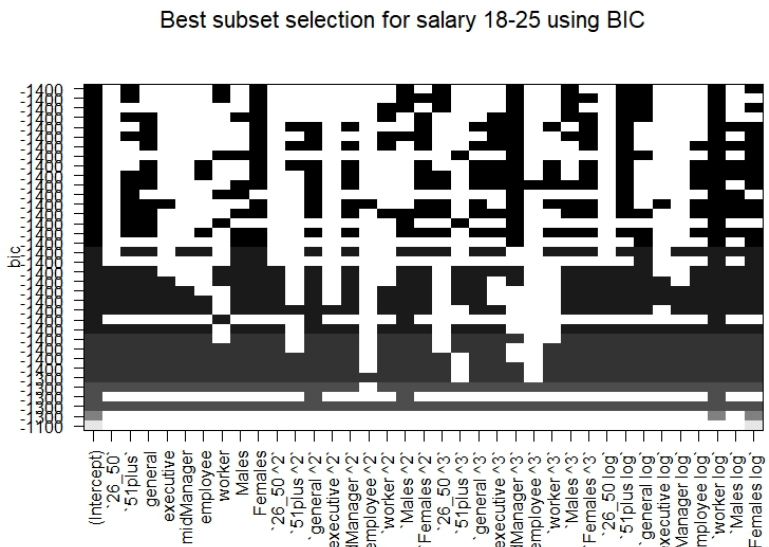


Prediction for young people

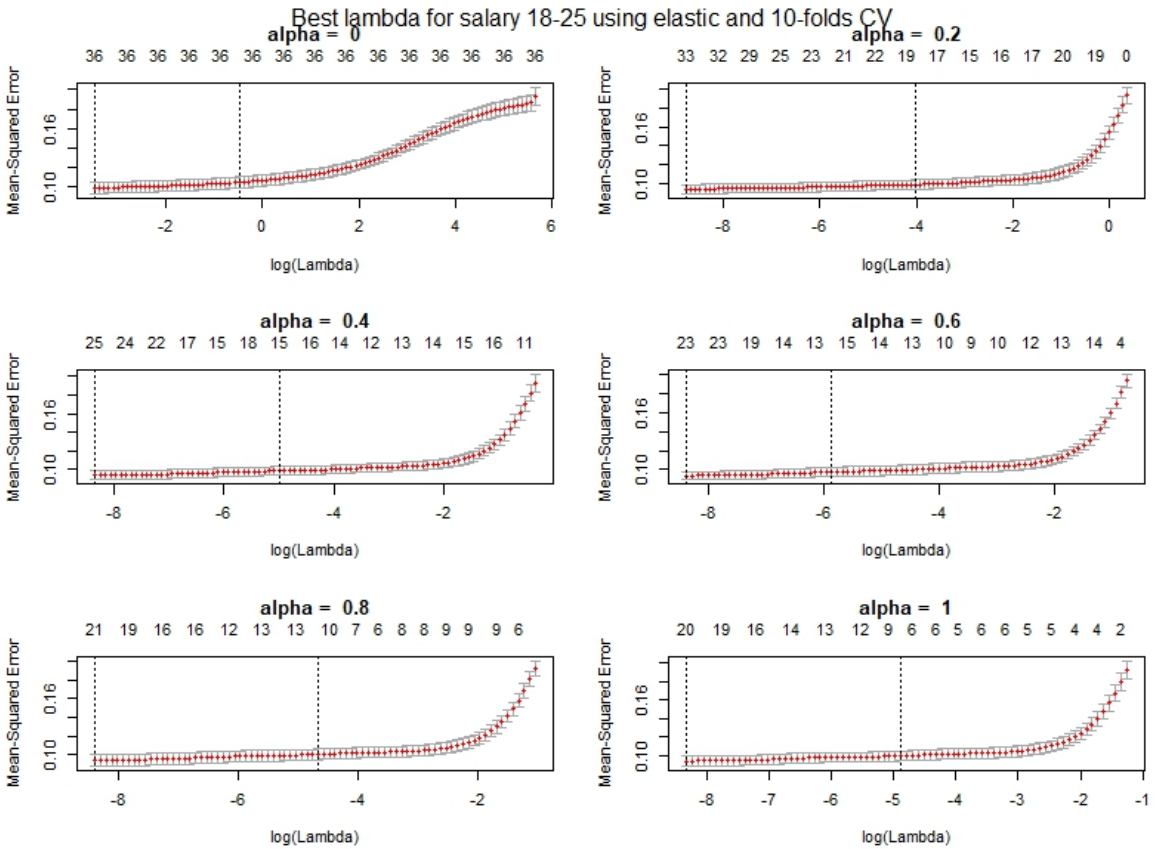




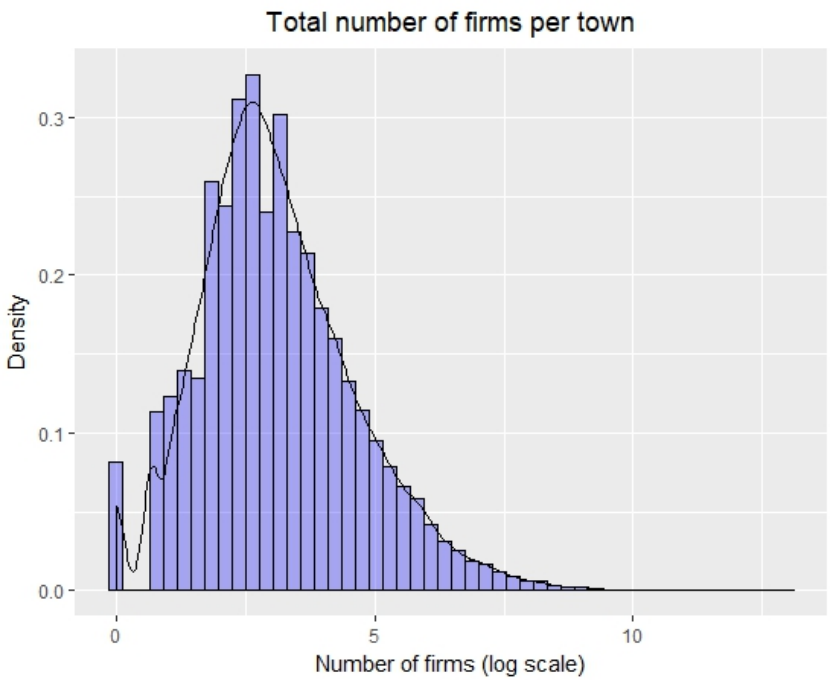
Prediction for young people using BSS

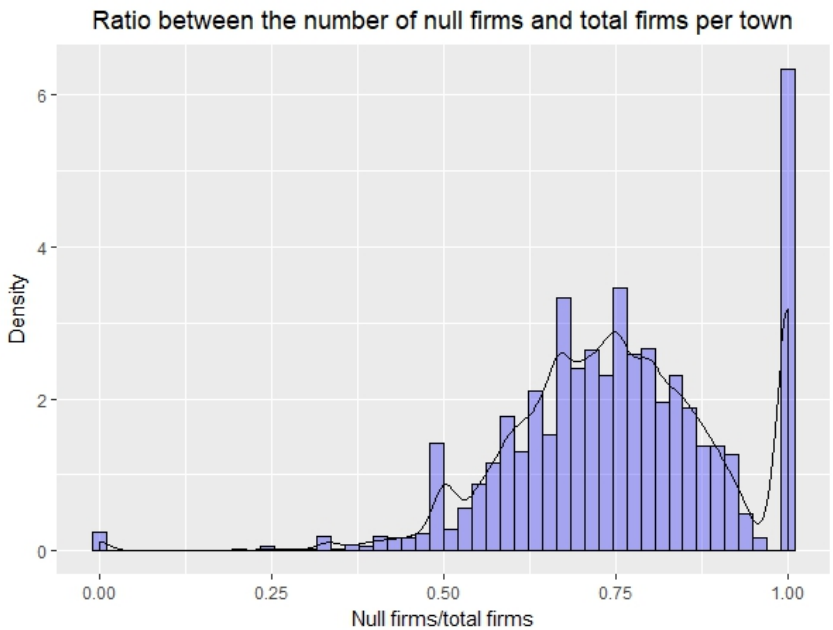


Elastic net and and 10-folds CV

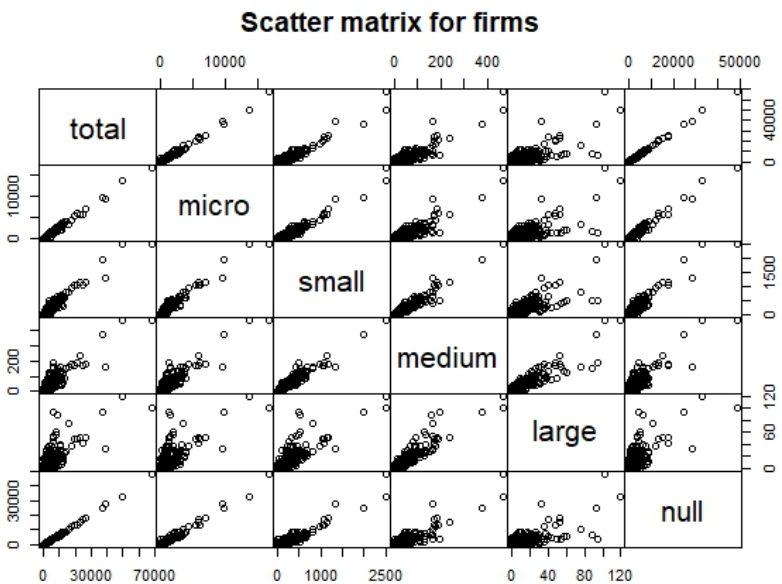


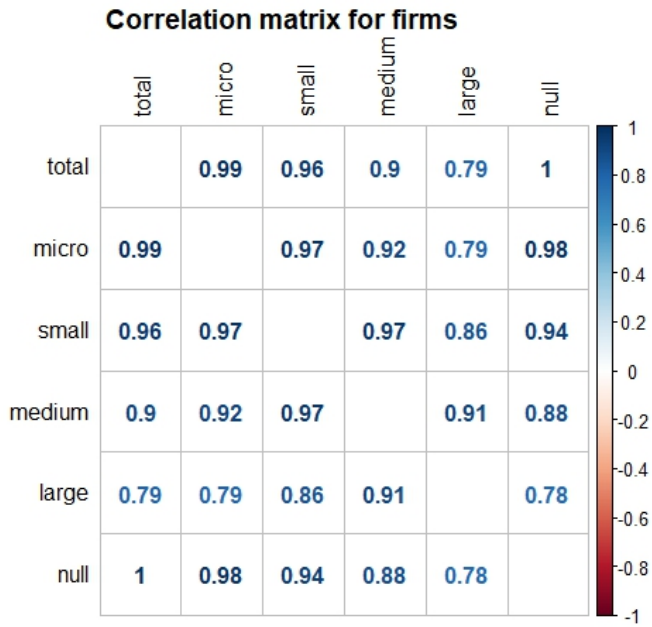
Distribution of firms per town





Bivariate relations

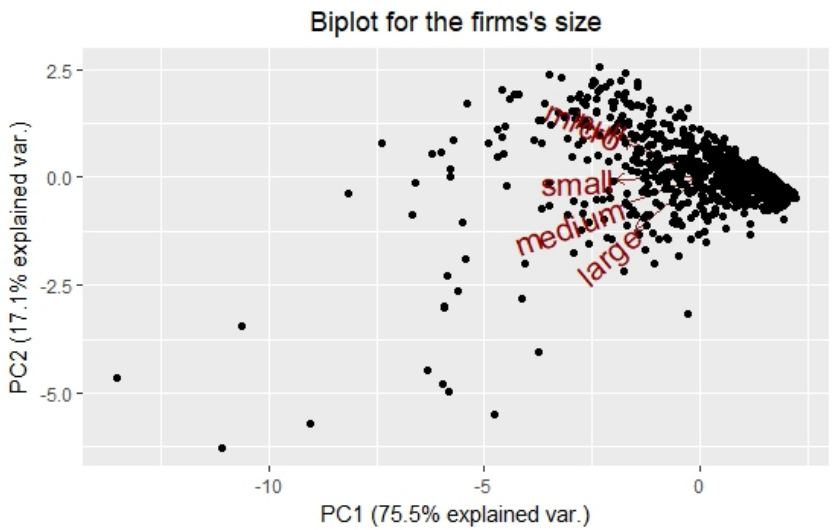


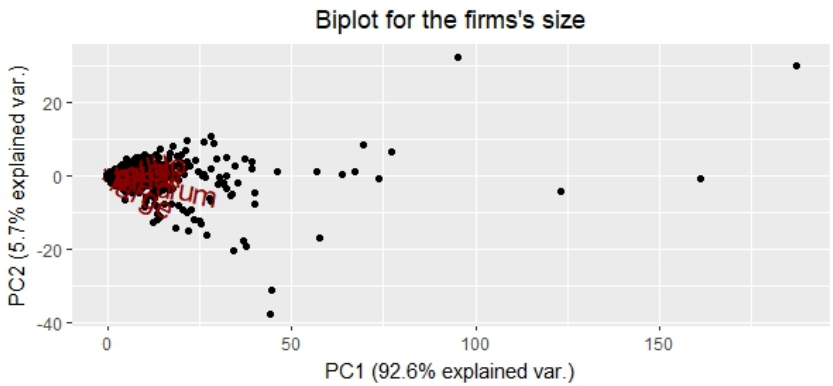


Excluding Paris

PCA

Using original data scaled (not logs)
Most typical vs. Excluding just Paris





Issues

- Unique code for salary data 1/7 of the total
- Loss of information when combining the separated datasets
- Missing additional information
- French DOM-TOM regions
- Outliers and spatial correlation

Future works

- Create meaningful indicators
- Clustering techniques to identify geographical clusters
- Verification and improvement of the obtained results
- Compare the methodologies used with robust ones
- Find complementary datasets

– Thank you –