

A study about salary difference in Brazil

Marcely Zanon Boito

December 21, 2016

The Datasets

In this report, I use datasets from the Brazilian Department of Labour, more specifically from the RAIS report (Social Informations Annual Report). These datasets contain information about all people registered as regular workers for the selected professions in 2014, following the “CBO” (brazilian official classification of professions), and the entries are from one of the 27 Brazil states.

This information is available because in Brazil, every time that an employer contracts, promotes or terminates an employee contract, it's mandatory to include this information in the government system. For this study, we have six datasets, each one representing a different profession: architecture, medicine, engineering, economy, law and street cleaning.

The Hypothesis:

Using this data, the objective is to identify how these different factors (age, gender, scholarly, profession, etc) can impact the average salary. More specifically, I would like to identify:

1. Is there a difference between the average salary between genders? If it is the case, in which profession we have the biggest salary gap per gender?
2. What is the impact that scholarly have in the average salary?
3. How does the age affect the salary?

Descriptives

We have seven variables in each dataset: scholarly (years), age (years), contract hours (hours per week), employment time (months), minimum salary (salary compared to the minimum wage) and average salary (brazilian reais). The table below was generated collecting the R “summary” command output for each profession. Each entry also has information about the gender of the employee, but since this information is categorical, it was omitted from the table.

- Number of observations:

Architect: 599

Civil Engineer: 2.239

Doctor: 4.214

Economist: 961

Lawyer: 2.476

Street Cleaner: 49.001

General Analysis for each variable:

- Scholarly:

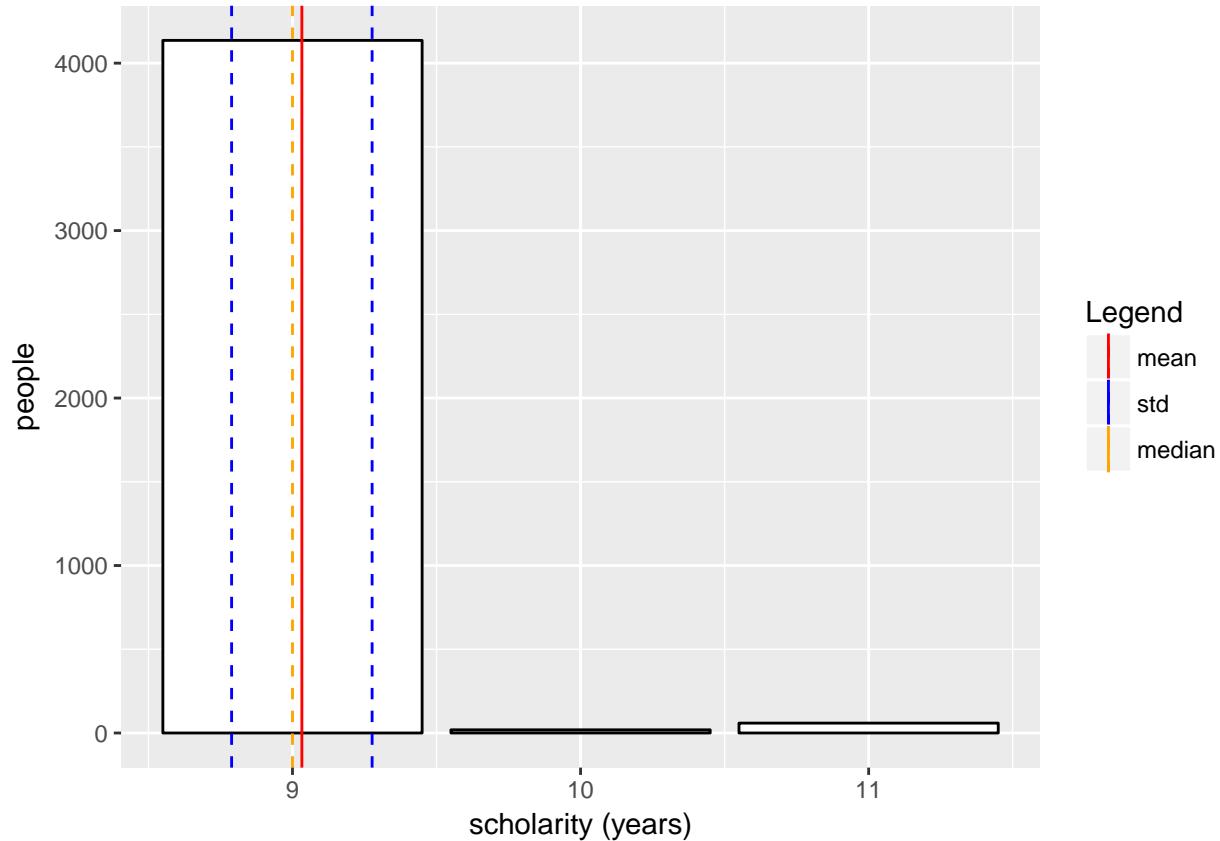
	Scholarity	Age	Contract hours	Employment time	Min salary	Avg salary
architect	Min.	5	17	5.00	0.60	272.50
	1 st Qu.	9	28	30.00	7.90	2,038.80
	Median	9	32	44.00	23.80	6.00
	Mean	9	35.76	37.91	59.51	7.30
	3 rd Qu.	9	43	44.00	58.85	8.97
	Max.	10	75	44.00	478.90	41.98
civil engineer	Min.	3	20	3	0.30	168.20
	1 st Qu.	9	29	35	10.50	6.00
	Median	9	34	40	27.90	8.41
	Mean	8.939	37.72	38.07	64.65	9.70
	3 rd Qu.	9	46	44	78.60	11.18
	Max.	11	88	44	484.40	93.93
doctor GP	Min.	9	23	1	0.20	0.32
	1 st Qu.	9	33	20	10.20	7.04
	Median	9	41	22	39.40	11.40
	Mean	9.033	43.12	27.26	91.09	12.09
	3 rd Qu.	9	54	40	134.80	16.81
	Max.	11	81	44	495.10	55.72
economist	Min.	5	18	8	0.40	0.58
	1 st Qu.	9	28	40	15.70	3.66
	Median	9	33	44	40.90	5.60
	Mean	9	36.19	41.75	89.17	8.11
	3 rd Qu.	9	44	44	101.90	9.28
	Max.	11	69	44	477.30	59.12
lawyer	Min.	4	19	1	0.00	0.30
	1 st Qu.	9	30	40	11.90	3.64
	Median	9	34	40	33.90	6.31
	Mean	9	37.17	38.13	69.72	8.54
	3 rd Qu.	9	43	44	78.33	10.53
	Max.	11	80	44	585.90	70.71
street cleaner	Min.	1	14	1	0.00	0.30
	1 st Qu.	4	32	40	6.90	1.11
	Median	5	41	44	21.40	1.35
	Mean	4.904	40.85	40.63	53.70	1.50
	3 rd Qu.	6	49	44	71.90	1.65
	Max.	11	92	44	542.80	15.81

Figure 1: Descriptives table

Inside professions, the values are really concentrated around the mean, and because of that it wouldn't be very helpful to analyse the impact of this variable inside a profession. However, since we have professions with a considerable distance between the means (e.g. doctor against street cleaner), we will try to compare how it impacts the salary.

Bellow we have an graphical example of how close the values are from the mean for this variable. The dataset used for this plot was the "doctor general practice".

```
library(ggplot2)
load(file="data/economist.Rdata")
load(file="data/street_cleaner.Rdata")
load(file="data/doctor_general_practice.Rdata")
load(file="data/architect.Rdata")
load(file="data/civil_engineer.Rdata")
load(file="data/lawyer.Rdata")
meanE <- mean(doctor_general_practice$Scholarity)
std <- sd(doctor_general_practice$Scholarity)
plot = ggplot(data = doctor_general_practice, aes(doctor_general_practice$Scholarity)) +
  geom_bar(fill="white", colour = "black") + labs(x= "scholarity (years)", y = "people") +
  geom_vline(aes(xintercept = meanE, colour = "mean")) +
  geom_vline(aes(xintercept = (meanE + std), colour = "std"), linetype = "dashed") +
  geom_vline(aes(xintercept = (meanE - std), colour = "std"), linetype = "dashed") +
  geom_vline(aes(xintercept = median(doctor_general_practice$Scholarity),
                 colour = "median"), linetype = "dashed") +
  scale_colour_manual(name = "Legend",
                      breaks = c("mean", "std", "median"),
                      values= c(mean = "red", std = "blue", median = "orange"))
plot
```

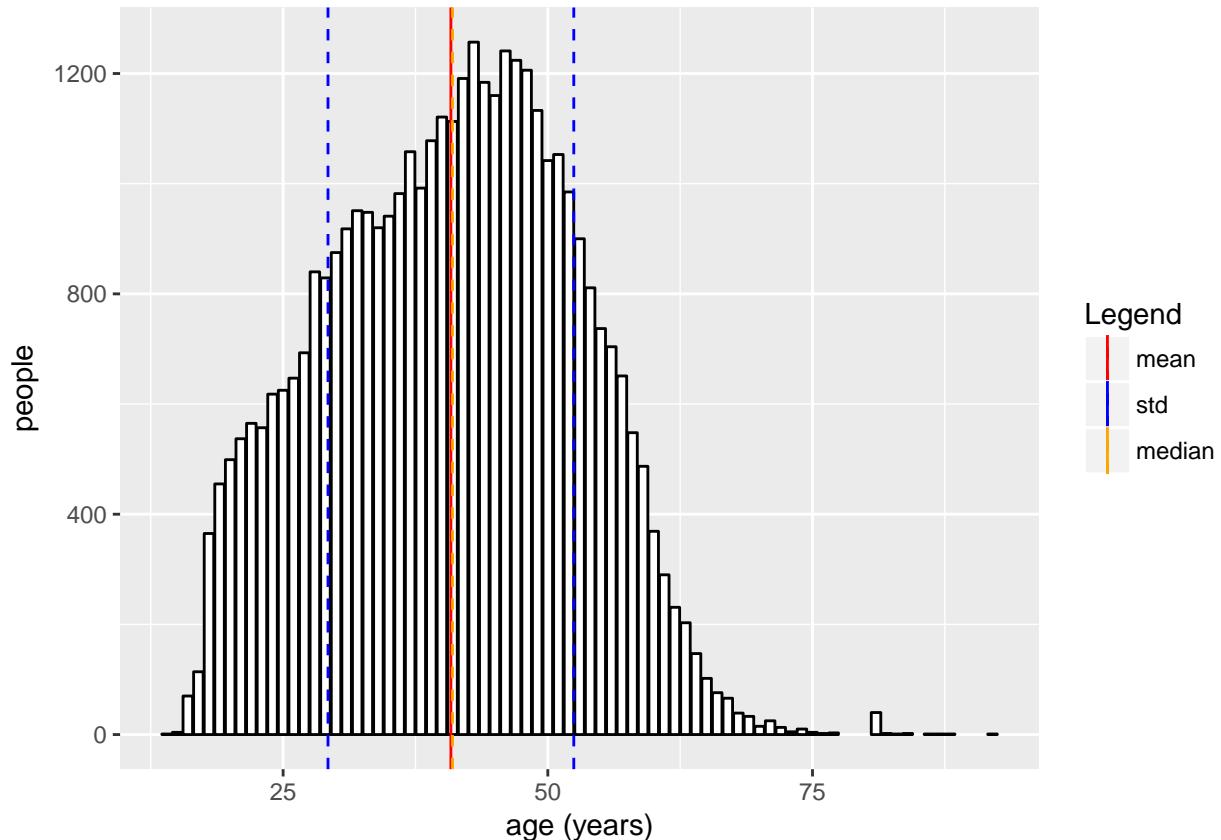


- Age:

In Brazil, it's possible to work after the 14 years (maximum of 6 hours per day until 16), and that's why we have observations for this age for the street cleaning dataset. However, for the other professions, since it's expected from the employee to have more time of education in order to fulfill their tasks, we have a higher minimum. For all the professions, we have a higher number of registers around 25 years, a common age to finish studies, and also around this age people have more mobility between jobs.

In Brazil, the age for retirement in 2014 was 59 years.

```
meanE <- mean(street_cleaner$age)
std <- sd(street_cleaner$age)
streetPlot = ggplot(data = street_cleaner, aes(street_cleaner$age)) +
  geom_bar(fill="white", colour = "black") + labs(x= "age (years)", y = "people") +
  geom_vline(aes(xintercept = meanE, colour = "mean")) +
  geom_vline(aes(xintercept = (meanE + std), colour = "std"), linetype = "dashed") +
  geom_vline(aes(xintercept = (meanE - std), colour = "std"), linetype = "dashed") +
  geom_vline(aes(xintercept = median(street_cleaner$age),
                 colour = "median"), linetype = "dashed") +
  scale_colour_manual(name = "Legend",
                      breaks = c("mean", "std", "median"),
                      values= c(mean = "red", std = "blue", median = "orange"))
streetPlot
```

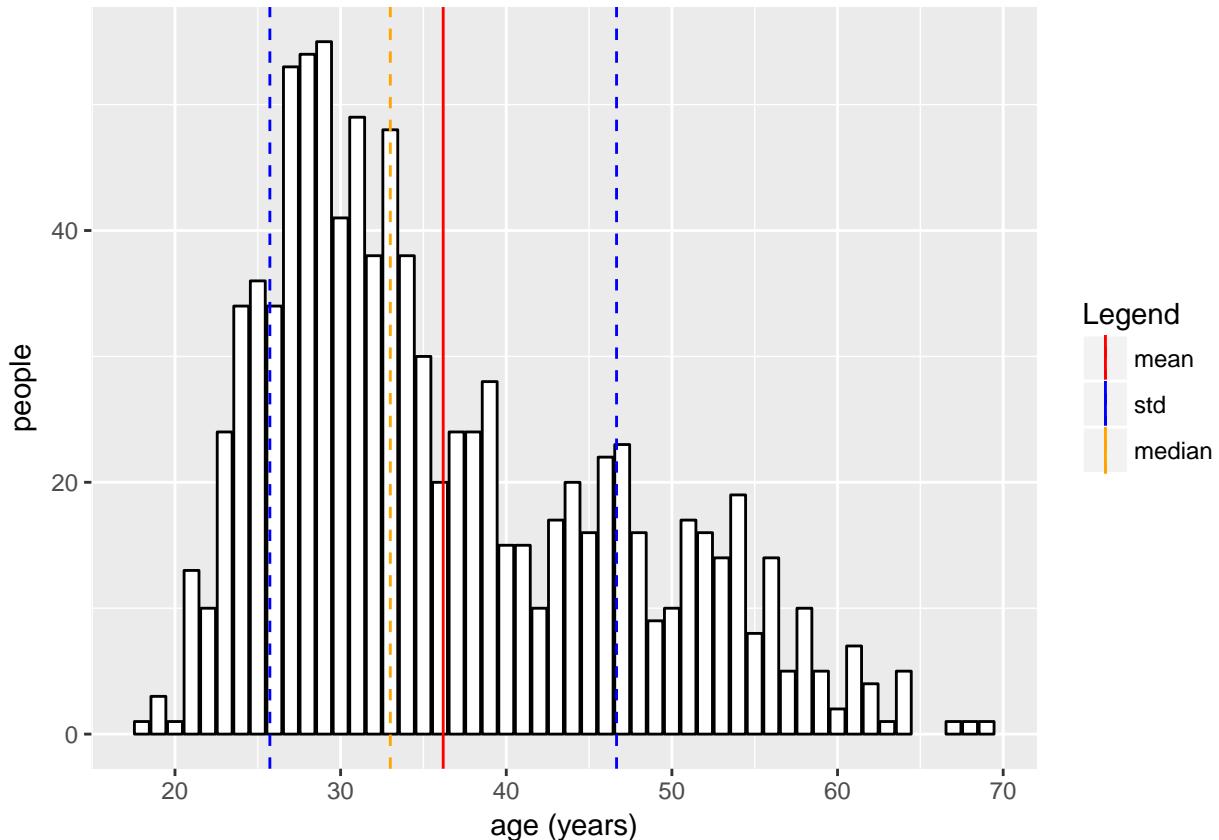


```
meanE <- mean(economist$age)
std <- sd(economist$age)
economistPlot = ggplot(data = economist, aes(economist$age)) +
  geom_bar(fill="white", colour = "black") + labs(x= "age (years)", y = "people")+
  geom_vline(aes(xintercept = meanE, colour = "mean")) +
```

```

geom_vline(aes(xintercept = (meanE + std), colour = "std"), linetype = "dashed") +
geom_vline(aes(xintercept = (meanE - std), colour = "std"), linetype = "dashed") +
geom_vline(aes(xintercept = median(economist$age),
               colour = "median"), linetype = "dashed") +
scale_colour_manual(name = "Legend",
                     breaks = c("mean", "std", "median"),
                     values = c(mean = "red", std = "blue", median = "orange"))
economistPlot

```



- **Contract Hours:**

In Brazil, the maximum number of hours per week allowed is 44, what explains the means concentrarating close to this number.

- **Gender:**

Below, a table showing the gender distribution in our dataset, separated by profession.

- **Employment Time:**

Employment time represents the quantity of months that the employee worked registered in a company. The graphs plotted below show the employment time per amount of people. The graphs show us that we have a greater amount of people with small employment time. Considering that in general we are supposed to have much more younger people working (smaller employment time) than seniors (bigger employment time), the graphs format make sense.

```

meanE <- mean(street_cleaner$employment_time)
std <- sd(street_cleaner$employment_time)
streetPlot = ggplot(data = street_cleaner, aes(street_cleaner$employment_time)) +
  geom_bar(fill="white", colour = "black") + labs(x= "employment Time (months)", y = "people") +

```

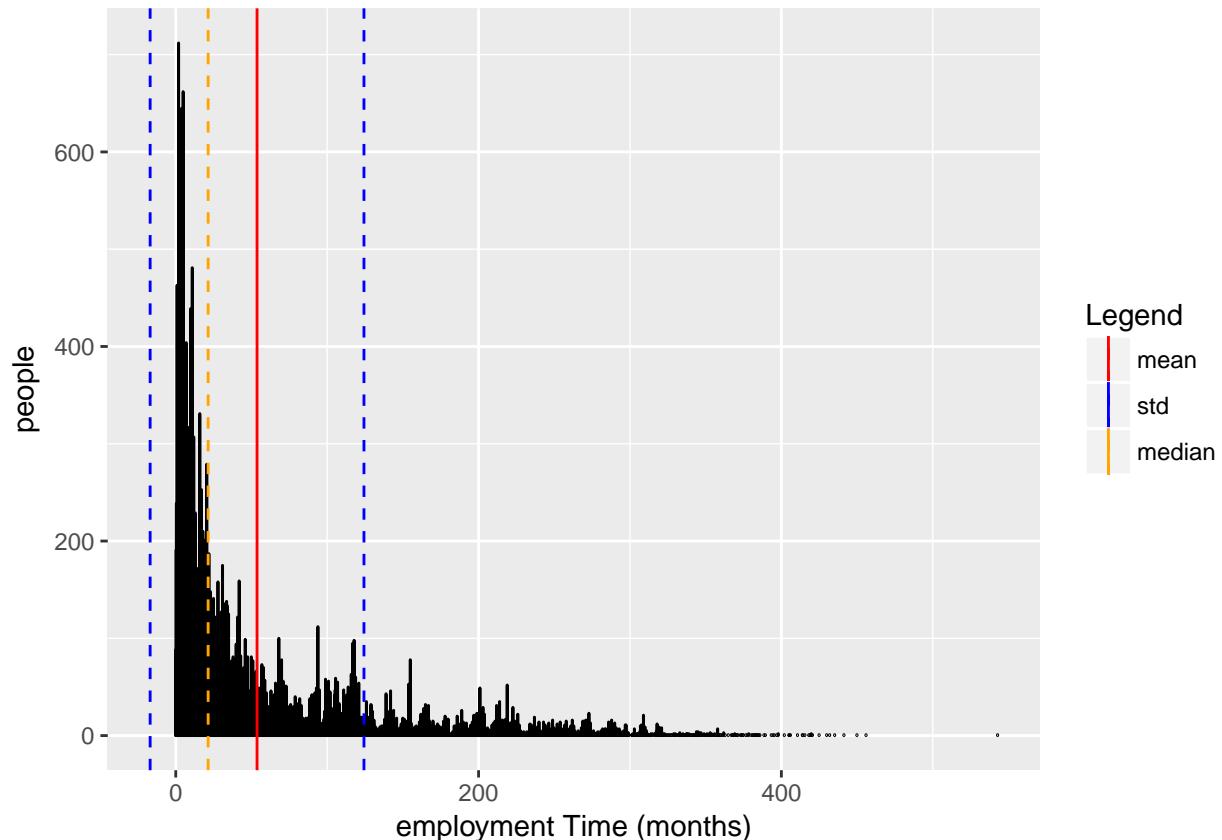
Profession	M	F
architect	47.25%	52.75%
civil engineer	79.86%	20.14%
doctor (GP)	67.11%	32.89%
economist	53.90%	46.10%
lawyer	53.39%	46.61%
street cleaner	33.47%	66.53%

Figure 2: Gender percentage per profession

```

geom_vline(aes(xintercept = meanE, colour = "mean")) +
geom_vline(aes(xintercept = (meanE + std), colour = "std"), linetype = "dashed") +
geom_vline(aes(xintercept = (meanE - std), colour = "std"), linetype = "dashed") +
geom_vline(aes(xintercept = median(street_cleaner$employment_time),
               colour = "median"), linetype = "dashed") +
scale_colour_manual(name = "Legend",
                     breaks = c("mean", "std", "median"),
                     values= c(mean = "red", std = "blue", median = "orange"))
streetPlot

```



```

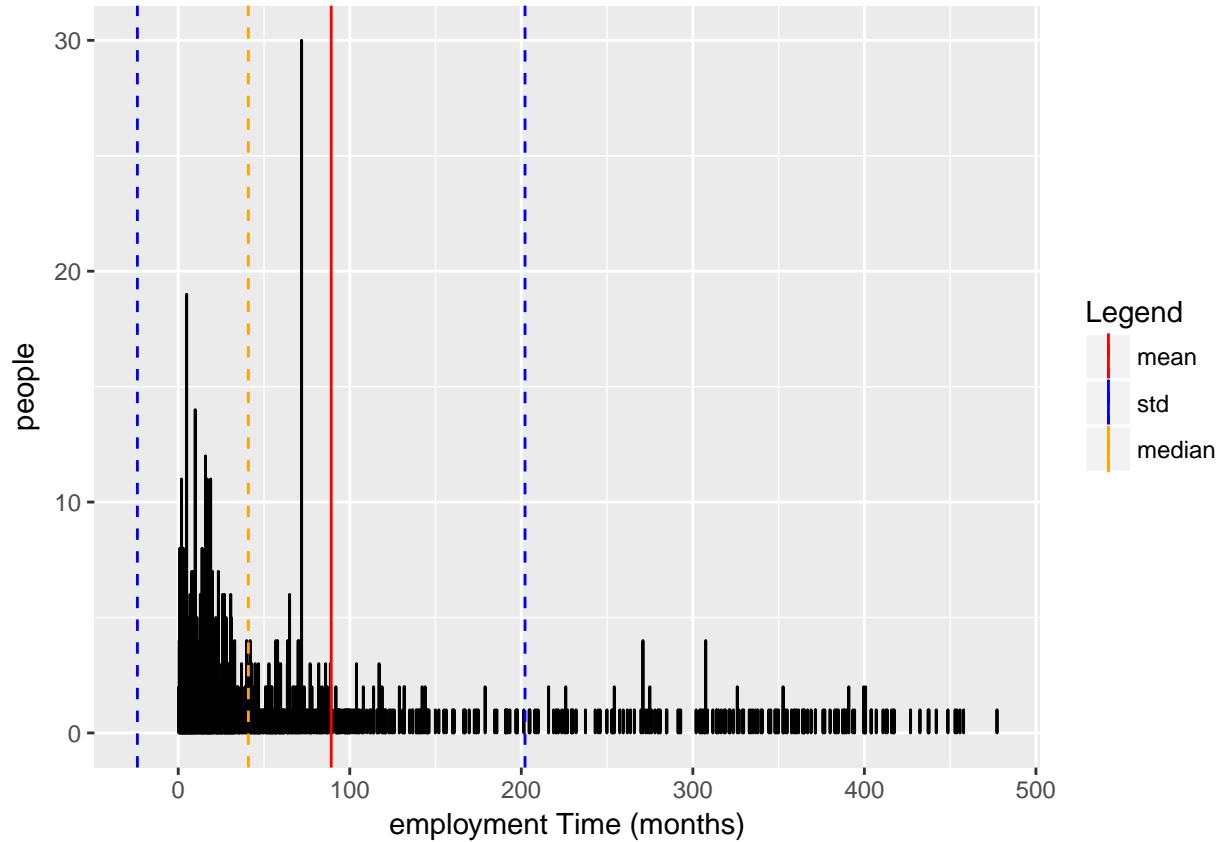
meanE <- mean(economist$employment_time)
std <- sd(economist$employment_time)

```

```

economistPlot = ggplot(data = economist, aes(economist$employment_time)) +
  geom_bar(fill="white", colour = "black") + labs(x= "employment Time (months)", y = "people")+
  geom_vline(aes(xintercept = meanE, colour = "mean")) +
  geom_vline(aes(xintercept = (meanE + std), colour = "std"), linetype = "dashed") +
  geom_vline(aes(xintercept = (meanE - std), colour = "std"), linetype = "dashed") +
  geom_vline(aes(xintercept = median(economist$employment_time),
                 colour = "median"), linetype = "dashed") +
  scale_colour_manual(name = "Legend",
                      breaks = c("mean", "std","median"),
                      values= c(mean = "red", std = "blue", median = "orange"))
economistPlot

```

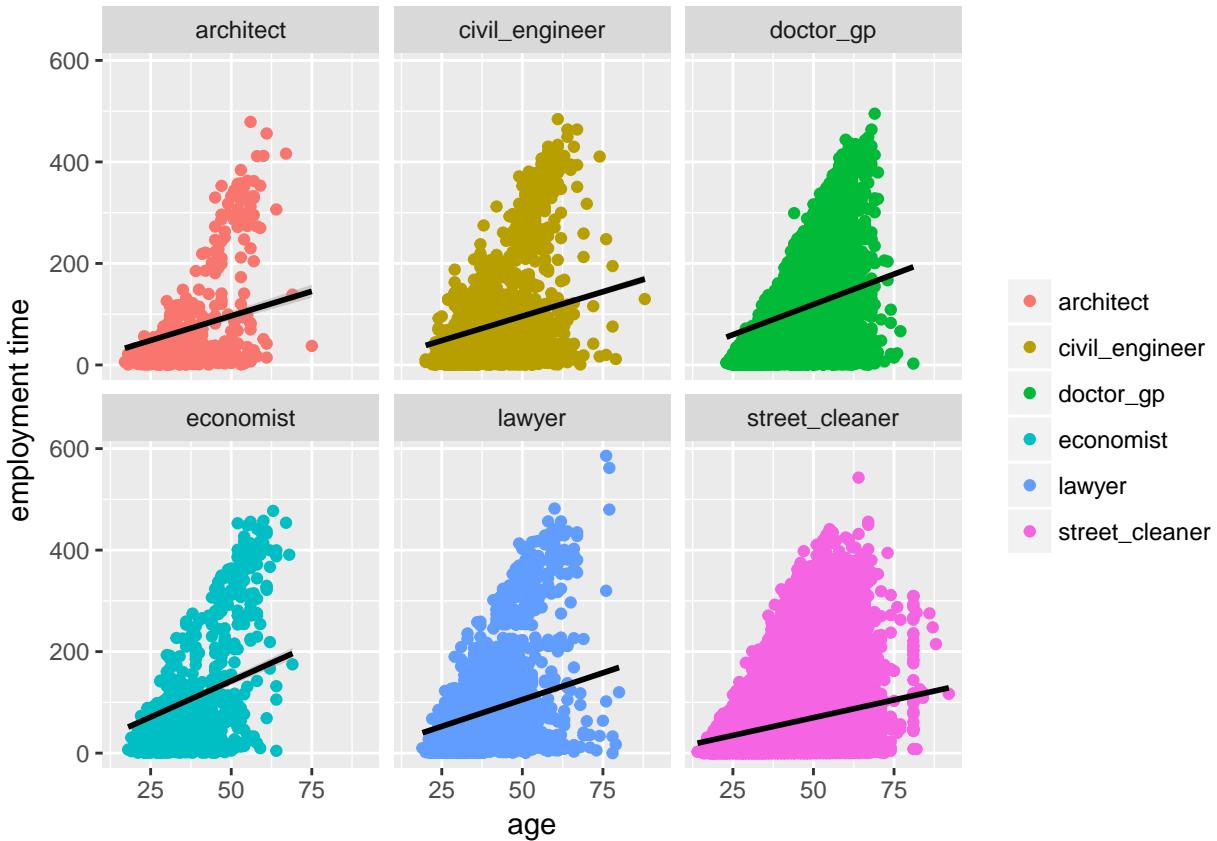


It's also interesting to explore the relationship between age and employment time. Does increasing the age mean increasing the employment time? Below I plotted these two, and tried to fit a linear model.

```

load(file="data/dataset.Rdata")
plot = ggplot(data = dataset, aes(dataset$age,dataset$employment_time, color = dataset$CB02002)) +
  geom_point() + labs(y = "employment time", x = "age") + geom_smooth(method= "lm",color= "black", form=
  facet_wrap(~ dataset$CB02002) + theme(legend.title=element_blank())
plot

```



In this model, I removed the intercept, since it would give unreasonable values for the variables in question (in no scenario a person with zero age would already have time of employment). Also, the graph above makes us think about the correlation between the variables, expressed below:

```
cor(dataset$employment_time, dataset$age)
```

```
## [1] 0.4403618
```

The variables are definitely positively correlated, but the correlation score is not that high. That and considering that we have a big cloud of points which can affect Pearson's correlation (since it is affected by the sample size), both variables were kept for the analysis.

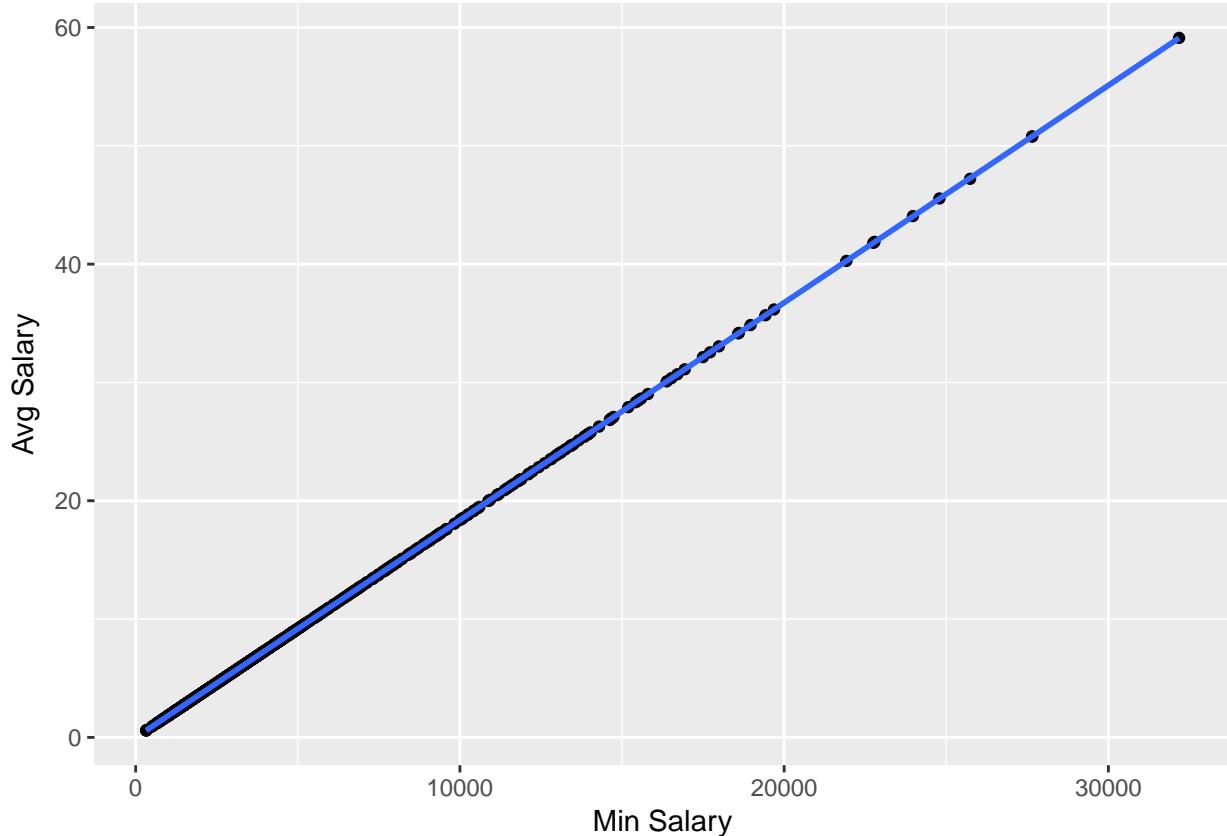
- **Min Salary:**

This variable represents the relation between the employee salary and the Brazilian minimum wage. In 2014, the national minimum wage was R\$ 724,00. However, this number is not absolute, since some sectors and states have different agreements for their minimum wages. Since it makes the comparison complicated (and to retrieve this information is equally difficult) we will consider the national minimum wage as reference for the analysis here.

The plot below is just an illustration about how the minimum salary is just a different way of viewing the average salary. Because of that, it wouldn't make sense to include it in our set of possible factors, and this variable will be excluded from further analysis in this report.

```
plot = ggplot(data = economist, aes(economist$avg_salary, economist$min_salary)) +
  geom_point() + labs(y = "Avg Salary", x = "Min Salary") + geom_smooth()
plot

## `geom_smooth()` using method = 'loess'
```

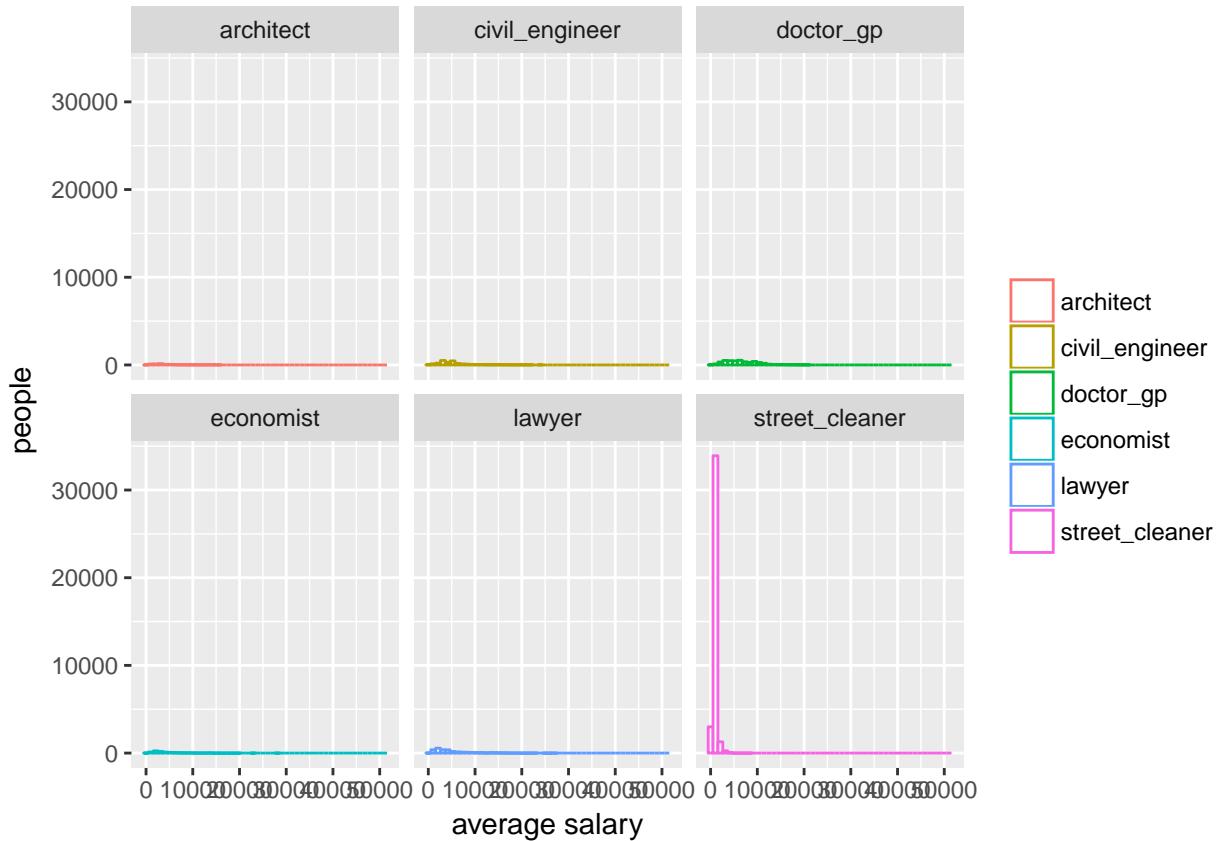


- **Average Salary:**

The variable whose behaviour we wish to summarize in a model. In the next section we explore its relationship with the other variables of the datasets.

```
# geom_point() + labs(y = "employment time", x = "age") + geom_smooth(method= "lm", color= "black", fo
# facet_wrap(~ dataset$CB02002)

plot = ggplot(data = dataset, aes(dataset$avg_salary, color = dataset$CB02002)) +
  geom_histogram(fill = "white", bins = 50) + labs(x= "average salary", y = "people") +
  facet_wrap(~ dataset$CB02002) + theme(legend.title=element_blank())
plot
```



This plot above is an histogram of the average salary per profession, and it does not give us a lot of information. It is difficult to have a suitable bucket number while not changing the visualization. Therefore, in the next section I start discussing how to scale this variable, in order to simplify the visualization and help the analysis.

Variables of interest against Average Salary

Before jumping to comparisons between salary and other variables, I normalize the average salary to allow a fair comparison between the observations in our datasets. This normalization is needed because without it we could conclude that some cases in our dataset are receiving more or less money because of a given factor, where in fact it could only happen that these people are working more than the others! Therefore, from now on I'm going to use the average salary divided by the number of hours worked, and I'm going to refer to this value as "normalized average salary".

A second transformation that I do in our target variable is not related to its meaning, but just a range transformation. In order to better visualize and summarize the normalized average salary, for this analysis we are going to apply a natural logarithm on it, since it does not change the values order, just scales it.

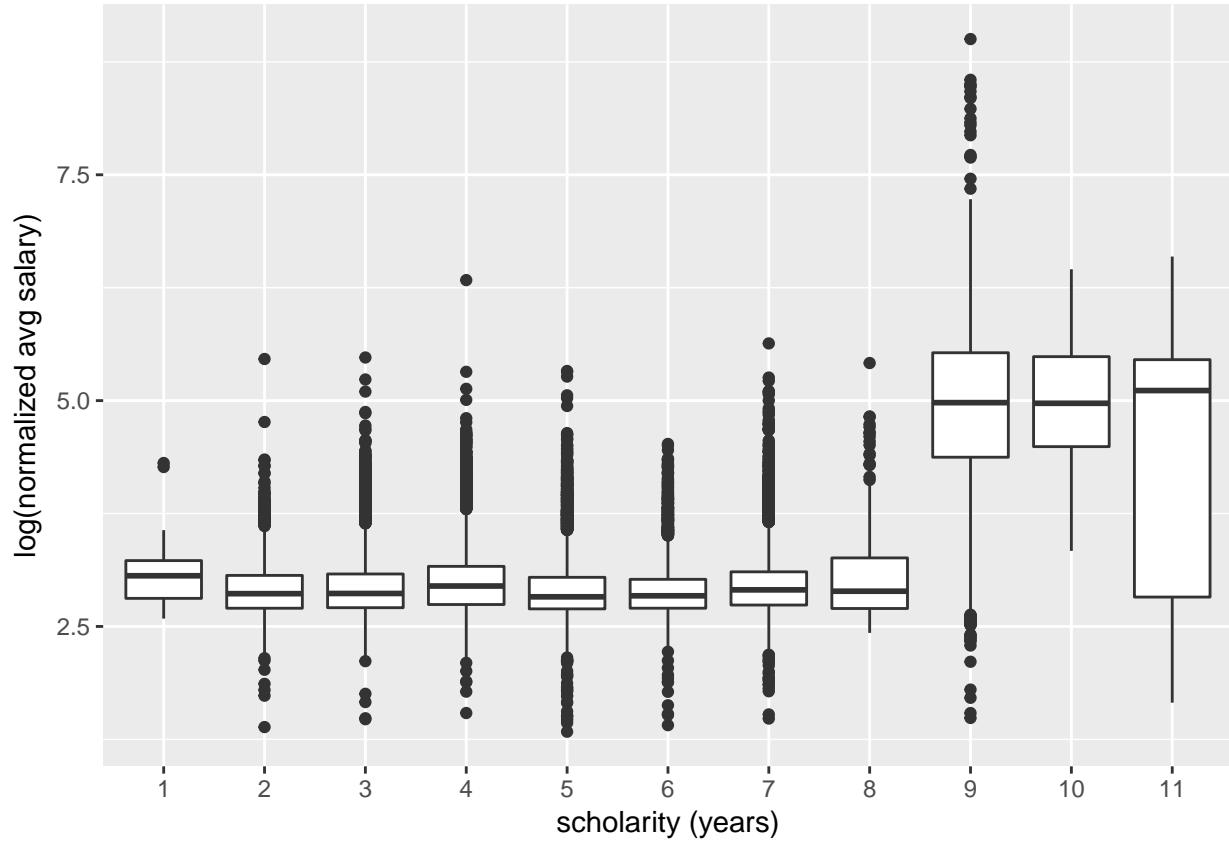
It is important to consider that applying the log operation on the average salary will change the interpretation of the results for the final regression, and it will be interpreted as the percentage salary variance given the absolute variance for the other variables.

Scholarity

As discussed before, for this comparison, is not very useful to look at the discrepancies inside a profession, and because of that, here we compare the ensemble of professions based solely on their normalized average

salary and their scholarly.

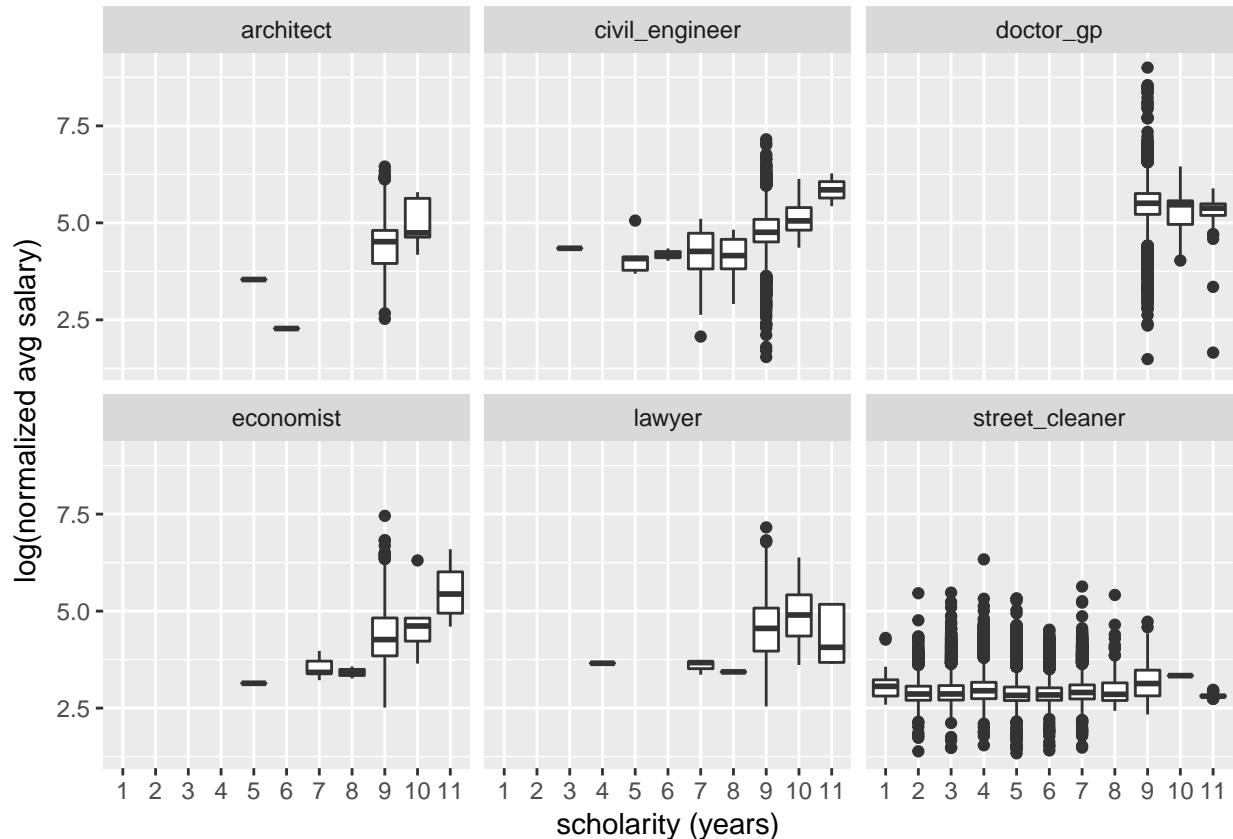
```
plot = ggplot(data= dataset, aes(factor(dataset$Scholarity), log(dataset$norm_avg_salary))) +  
  geom_boxplot() + labs(x = "scholarity (years)", y = "log(normalized avg salary)")  
plot
```



This visualization help us to identify a possible relationship between the salary and the scholarly. Although between the first seven years of scholarly we do not seem to have a considerable difference, from the eighth to the ninth year, it seems to impact drastically the average salary of the employee.

When plotting this, the first thing that I thought was: but, it is fair? My fear was that I would be comparing two disjoint sets: a set that goes from one year of scholarly to eight, and a different set of professions that require at least nine years of scholarly. However, fortunately, after plotting scholarly per profession (below) I was able to see that, even if we do have a profession where it seems to be a requirement to be above eight (medicine), all the others professions overlap, what makes the comparison valid.

```
plot = ggplot(data= dataset, aes(factor(dataset$Scholarity), log(dataset$norm_avg_salary))) +  
  geom_boxplot() + labs(x = "scholarity (years)", y = "log(normalized avg salary)") +  
  facet_wrap(~ dataset$CB02002)  
plot
```



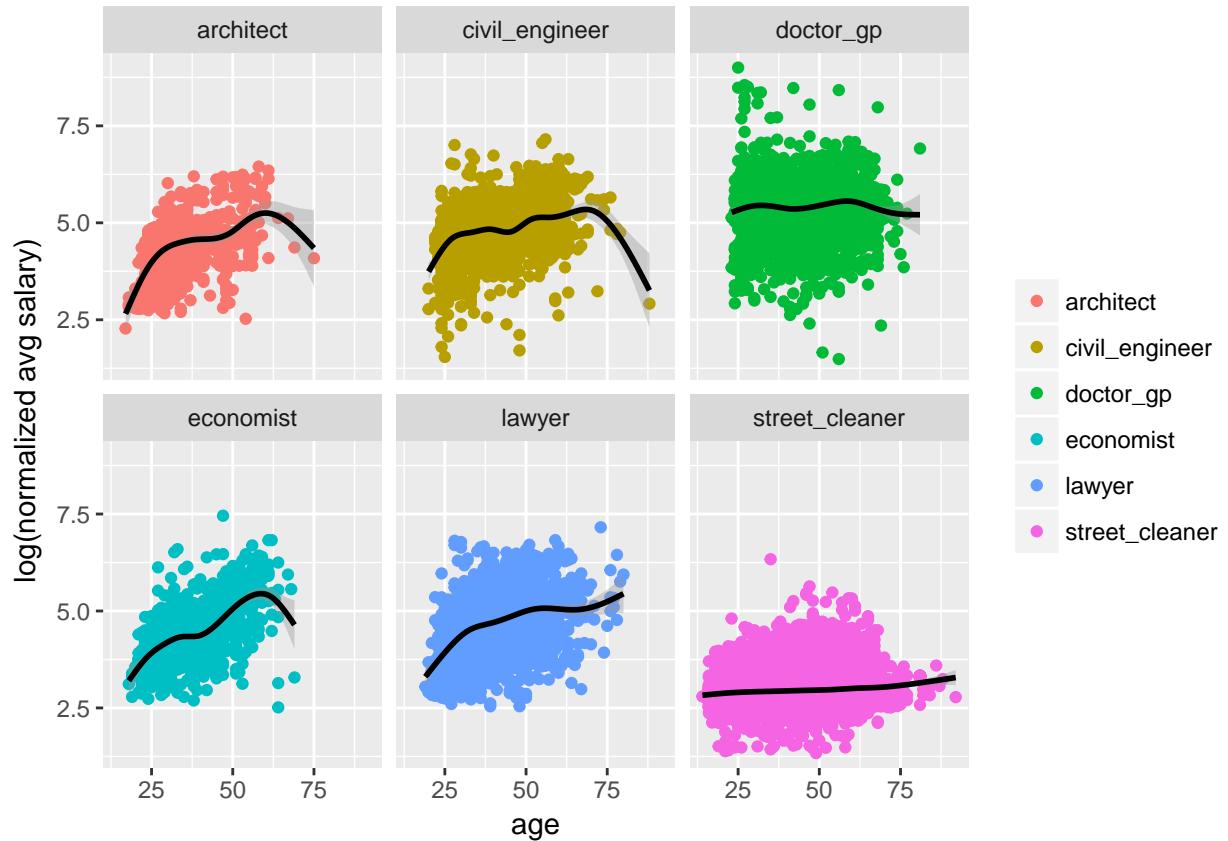
Age

```

plot = ggplot(data = dataset, aes(dataset$age, log(dataset$norm_avg_salary),
                                 color = dataset$CB02002)) + geom_point() +
  geom_smooth(color = "black") + labs(x = "age", y = "log(normalized avg salary)") +
  facet_wrap(~ dataset$CB02002) + theme(legend.title=element_blank())
plot

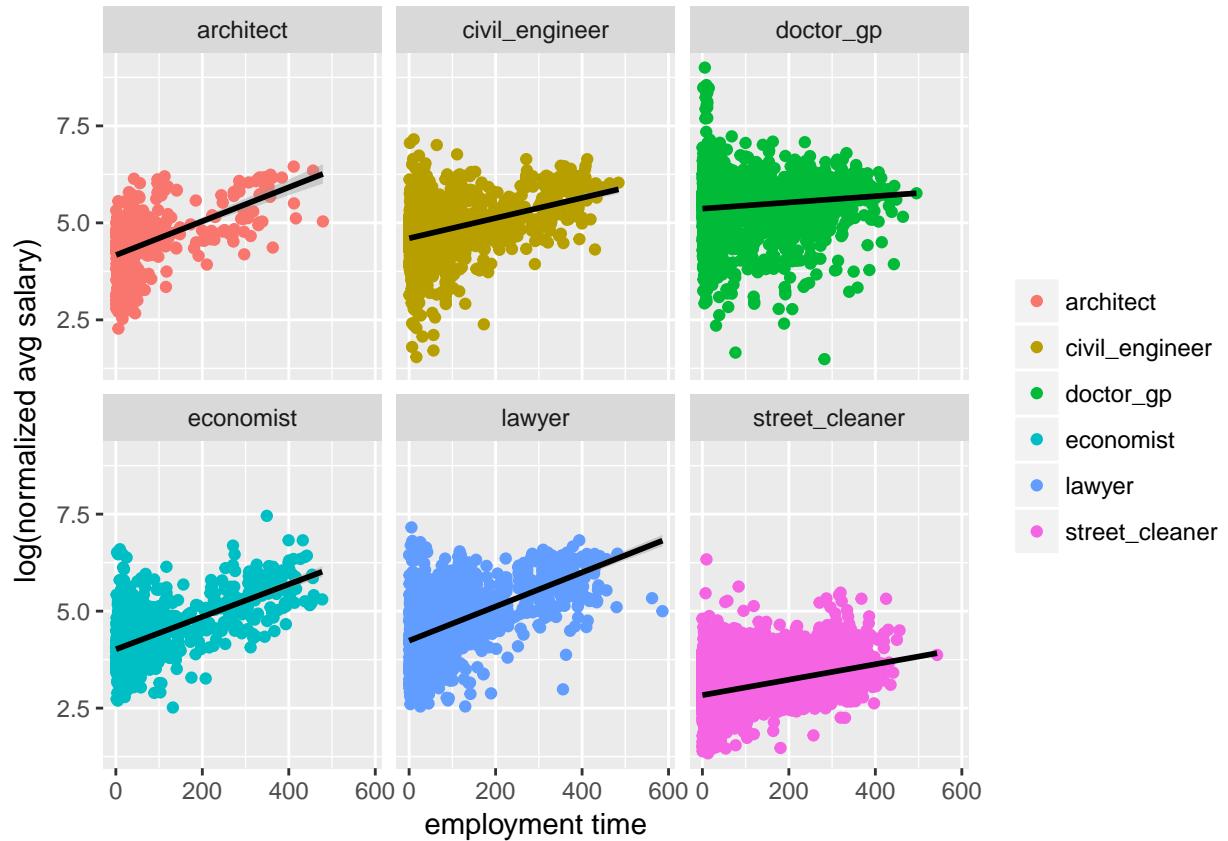
## `geom_smooth()` using method = 'gam'

```



Employment Time

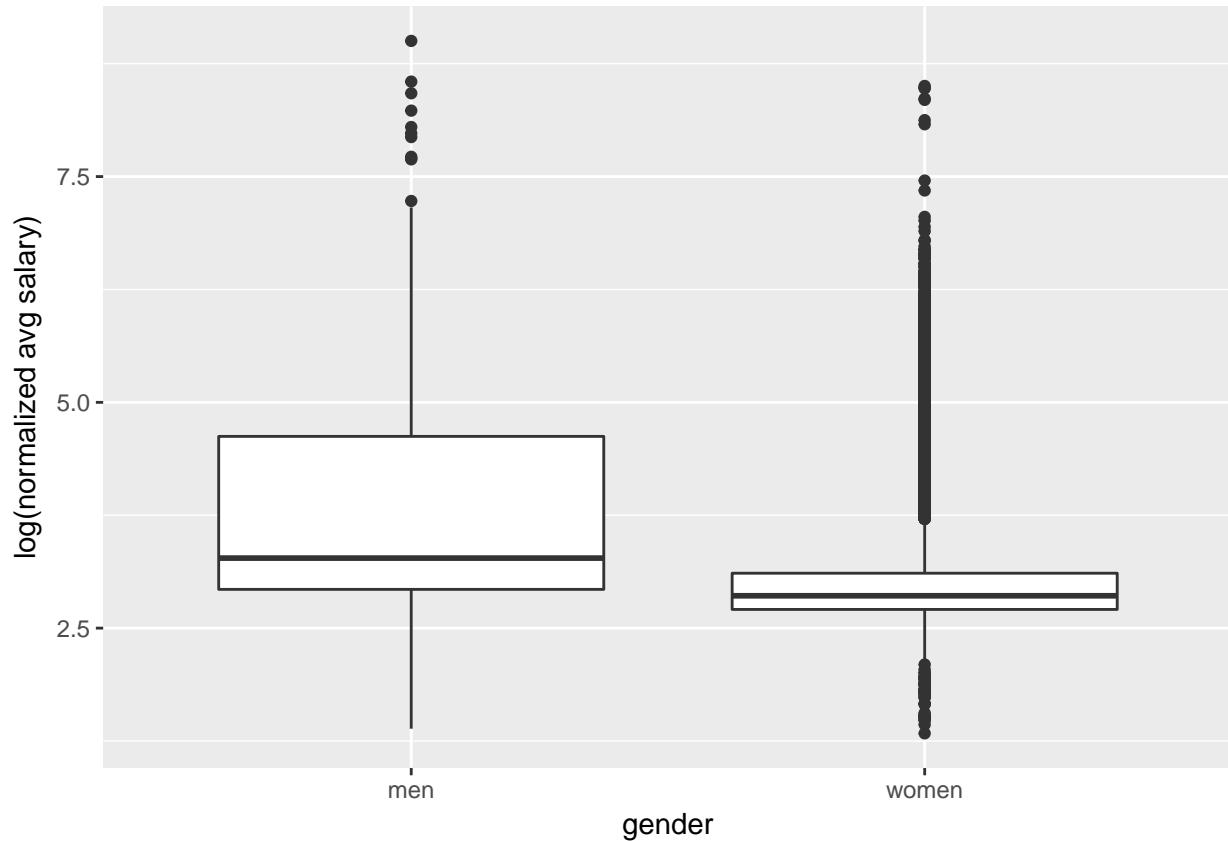
```
plot = ggplot(data = dataset, aes(dataset$employment_time, log(dataset$norm_avg_salary),
                                    color = dataset$CB02002)) + geom_point() +
  geom_smooth(color = "black", method = lm) + labs(x = "employment time",
                                                 y = "log(normalized avg salary)") +
  facet_wrap(~ dataset$CB02002) + theme(legend.title=element_blank())
plot
```



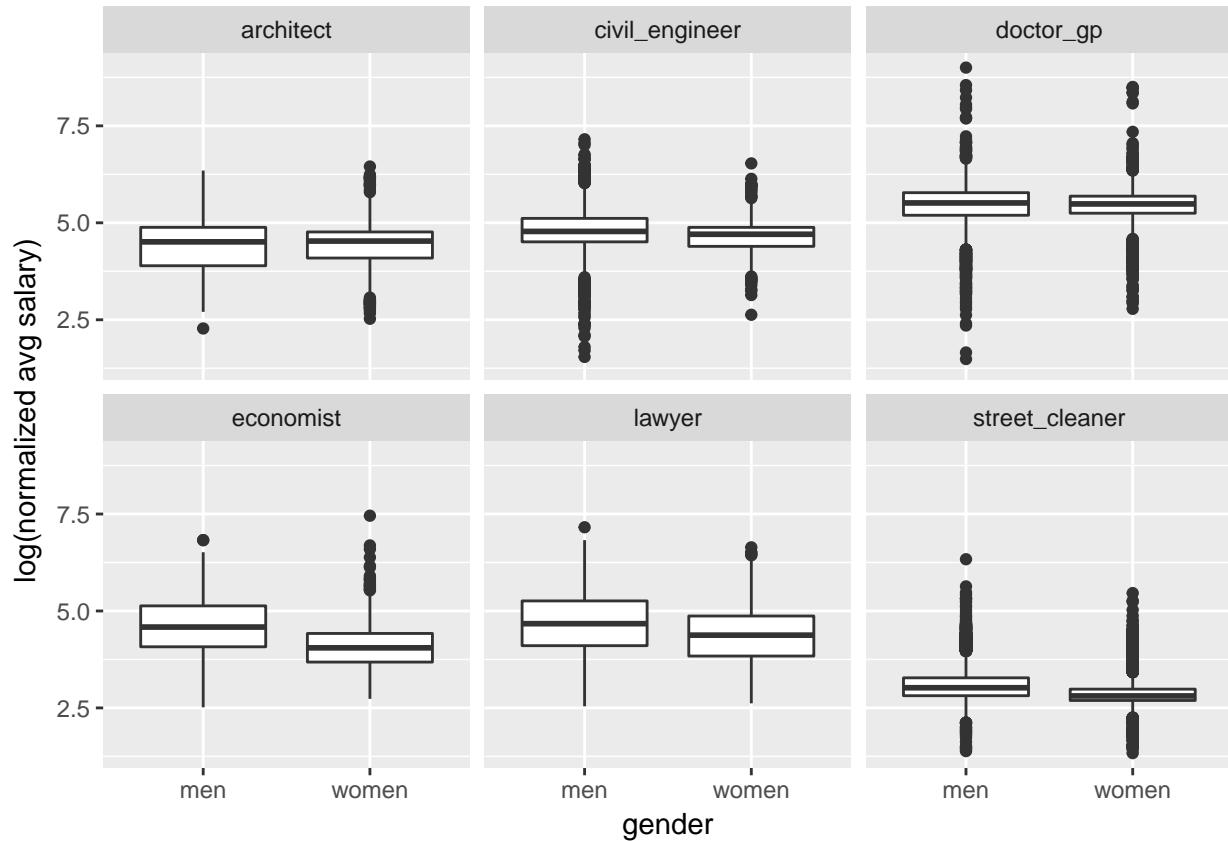
Similar as when comparing age with employment time, the observations seem to have a huge variability.

Gender

```
plot = ggplot(data=dataset, aes(x= factor(dataset$gender, labels=c("men","women")),
                                 y=log(dataset$norm_avg_salary))) +
  geom_boxplot() + labs(x = "gender", y = "log(normalized avg salary)")
plot
```



```
plot = ggplot(data=dataset, aes(x= factor(dataset$gender, labels=c("men","women")),
                                 y=log(dataset$norm_avg_salary))) +
  geom_boxplot() + labs(x = "gender", y = "log(normalized avg salary)") +
  facet_wrap(~ dataset$CB02002)
plot
```



Mincer Earnings Function

The Mincer earnings function was published in 1974 and it models the salary in function of years of education (scholarity), experience in the market (here approximated by the age) and other factors (such as gender and employment time). It is one of the most widely used models in empirical economics, and more information can be found at [1].

The Mincer model used in this report is the combination of all the variables investigated in the last section. It is defined as follows:

$$\ln \text{norm_avg_salary} = B_0 + B_1 \text{scholarity} + B_2 \text{age} + B_3 \text{age}^2 + B_4 \text{gender} + B_5 \text{employment_time} + \text{error}$$

Where B_0 is the salary of an individual with no education and no experience, and B_1, B_2, B_3, B_4 and B_5 are the adjust parameters.

```

reg = lm(data = dataset,
          formula = log(dataset$norm_avg_salary) ~ dataset$Scholarity +
            dataset$age + dataset$gender + dataset$employment_time +
            I(dataset$age^2))

summary(reg)

##
## Call:
## lm(formula = log(dataset$norm_avg_salary) ~ dataset$Scholarity +
##     dataset$age + dataset$gender + dataset$employment_time +
##     I(dataset$age^2), data = dataset)

```

```

## 
## Residuals:
##   Min     1Q Median     3Q    Max 
## -3.7235 -0.4319 -0.0591  0.3628  4.8272 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           1.442e+00  3.216e-02 44.82   <2e-16 *** 
## dataset$Scholarity   2.699e-01  1.307e-03 206.46   <2e-16 *** 
## dataset$age            3.694e-02  1.520e-03  24.30   <2e-16 *** 
## dataset$gender        -4.024e-01  5.934e-03 -67.82   <2e-16 *** 
## dataset$employment_time 3.002e-03  3.983e-05  75.37   <2e-16 *** 
## I(dataset$age^2)      -3.743e-04  1.812e-05 -20.65   <2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 0.6211 on 48995 degrees of freedom 
## Multiple R-squared:  0.5674, Adjusted R-squared:  0.5674 
## F-statistic: 1.285e+04 on 5 and 48995 DF, p-value: < 2.2e-16

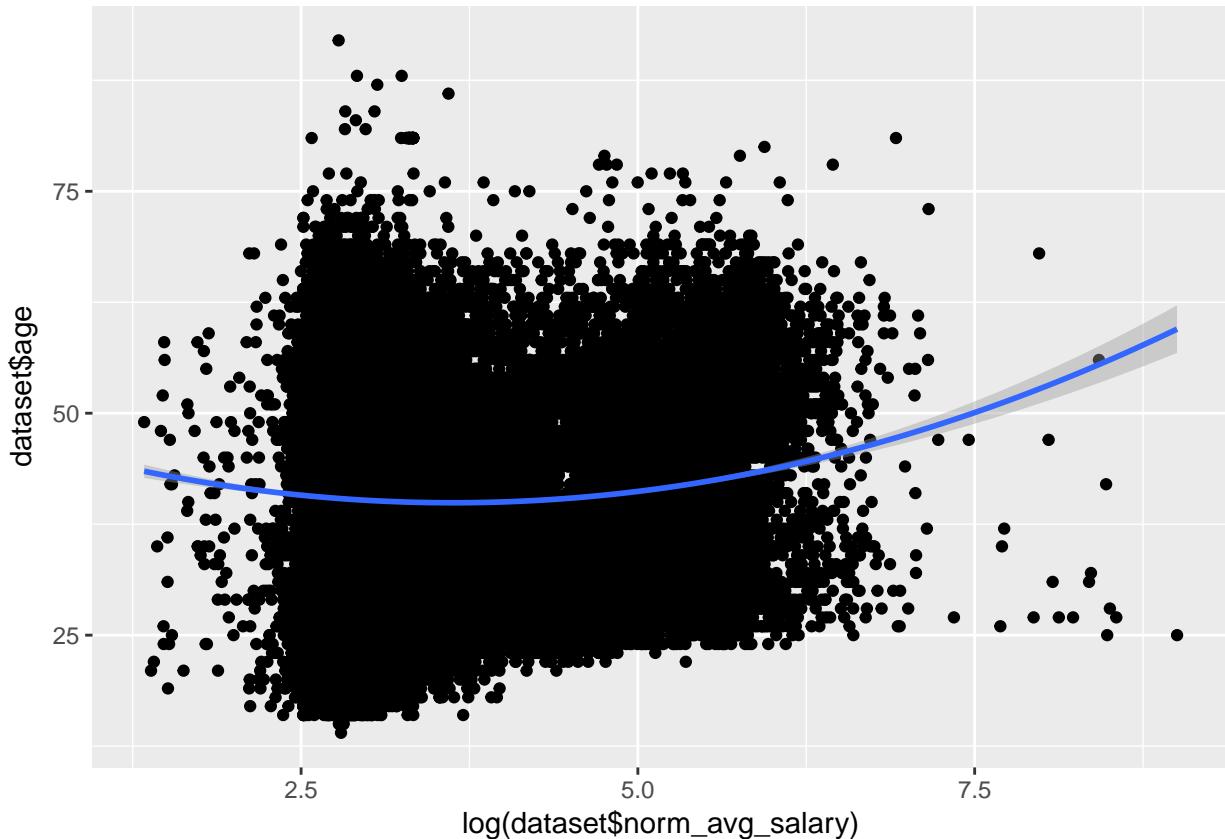
```

The regression summary confirms the relationship between our factor and response variables, discarding totally the null hypothesis. However, we do not have a good adjusted R-squared value (only 0.5674), what probably can be explained by the noisy nature of our dataset, thing that was illustrated in the last section and it is again shown in the plot bellow:

```

plot = ggplot(data = dataset, aes(log(dataset$norm_avg_salary),dataset$age)) + 
  geom_point() + geom_smooth(method = lm, formula = y ~ x + I(x^2))
plot

```



With this level of noise, I would say it is impossible to fit a model with a very good R-squared value!

Also, a pitfall that I fell when I was trying to fit this regression was to remove the intercept for this model, which give us the summary bellow:

```
reg = lm(data = dataset, formula = log(dataset$norm_avg_salary) ~ 0 +
          dataset$Scholarity + dataset$age + dataset$gender +
          dataset$employment_time + I(dataset$age^2))
summary(reg)

##
## Call:
## lm(formula = log(dataset$norm_avg_salary) ~ 0 + dataset$Scholarity +
##     dataset$age + dataset$gender + dataset$employment_time +
##     I(dataset$age^2), data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.7460 -0.4217 -0.0253  0.4017  4.9536
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## dataset$Scholarity        2.902e-01  1.251e-03 231.95 <2e-16 ***
## dataset$age                 9.725e-02  7.213e-04 134.83 <2e-16 ***
## dataset$gender              -3.536e-01  5.951e-03 -59.41 <2e-16 ***
## dataset$employment_time    2.939e-03  4.062e-05  72.36 <2e-16 ***
## I(dataset$age^2)           -1.053e-03  1.017e-05 -103.47 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6337 on 48996 degrees of freedom
## Multiple R-squared:  0.9672, Adjusted R-squared:  0.9672
## F-statistic: 2.887e+05 on 5 and 48996 DF, p-value: < 2.2e-16
```

Looking just at the adjusted R-squared, it looks a perfect fit. However, removing the intercept (B_0 at the mincer model defined above) is not right for two reasons. First because it is part of the model, and therefore it has a weight on it. The second reason I discovered after some research, looking at the discussion at [2].

Basically, what happens when one removes the intercept is that R uses a modified form of the equation to compare the current model to the reference model (that only contains the intercept). Since it does not have the intercept, it implicitly uses as reference model a model made of noise only. In cases as this dataset, where we have a huge cloud of points, this factor ends up dominating the value, giving us the false impression of fitting.

Finally, a final hint that removing the intercept is not a good idea, is our number of degrees of freedom, which is ridiculously small compared to the size of our dataset (49.001 observations).

Looking at each profession

In the regression that was made, we considered the whole dataset. This is problematic because we are assuming that the relationship is the same between these factors and the response variable.

For example, maybe age for street cleaning is a factor that affects the salary greatly, lowering it, while for professions like medicine, it could mean that this doctor is more experienced and therefore the salary would be better. Therefore, in this section we explore the different relationships for the different professions.

Architect

```
reg = lm(data = architect, formula = log(architect$norm_avg_salary) ~ +
          architect$Scholarity + architect$age + architect$gender +
          architect$employment_time + I(architect$age^2))
summary(reg)

##
## Call:
## lm(formula = log(architect$norm_avg_salary) ~ +architect$Scholarity +
##      architect$age + architect$gender + architect$employment_time +
##      I(architect$age^2), data = architect)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -2.00402 -0.40604  0.08274  0.43859  1.71639
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             -0.9584929  0.9854734 -0.973  0.33114  
## architect$Scholarity    0.2816848  0.1072738  2.626  0.00887 ** 
## architect$age            0.1168276  0.0171731  6.803 2.51e-11 *** 
## architect$gender         0.1417396  0.0516645  2.743  0.00626 ** 
## architect$employment_time 0.0035737  0.0003490 10.239 < 2e-16 *** 
## I(architect$age^2)        -0.0012670  0.0002148 -5.899 6.16e-09 *** 
## ---                        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6106 on 593 degrees of freedom
## Multiple R-squared:  0.3574, Adjusted R-squared:  0.352 
## F-statistic: 65.98 on 5 and 593 DF,  p-value: < 2.2e-16
```

Civil Engineer

```
reg = lm(data = civil_engineer, formula = log(civil_engineer$norm_avg_salary) ~ +
          civil_engineer$Scholarity + civil_engineer$age + civil_engineer$gender +
          civil_engineer$employment_time + I(civil_engineer$age^2))
summary(reg)

##
## Call:
## lm(formula = log(civil_engineer$norm_avg_salary) ~ +civil_engineer$Scholarity +
##      civil_engineer$age + civil_engineer$gender + civil_engineer$employment_time +
##      I(civil_engineer$age^2), data = civil_engineer)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -3.2236 -0.2586  0.0411  0.3046  2.3324
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             1.807e+00  2.914e-01  6.203 6.60e-10 *** 
## civil_engineer$Scholarity 2.007e-01  2.849e-02  7.044 2.47e-12 ***
```

```

## civil_engineer$age           4.870e-02  7.254e-03   6.713 2.40e-11 ***
## civil_engineer$gender        -6.275e-02 2.977e-02  -2.108  0.0352 *
## civil_engineer$employment_time 1.895e-03  1.532e-04  12.367 < 2e-16 ***
## I(civil_engineer$age^2)      -4.607e-04 8.446e-05  -5.454 5.46e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5554 on 2233 degrees of freedom
## Multiple R-squared:  0.2022, Adjusted R-squared:  0.2004
## F-statistic: 113.2 on 5 and 2233 DF,  p-value: < 2.2e-16

```

Doctor General Practice

```

reg = lm(data = doctor_general_practice, formula = log(doctor_general_practice$norm_avg_salary) ~ +
          doctor_general_practice$Scholarity + doctor_general_practice$age + doctor_general_practice$gender +
          doctor_general_practice$employment_time + I(doctor_general_practice$age^2))
summary(reg)

##
## Call:
## lm(formula = log(doctor_general_practice$norm_avg_salary) ~ +doctor_general_practice$Scholarity +
##     doctor_general_practice$age + doctor_general_practice$gender +
##     doctor_general_practice$employment_time + I(doctor_general_practice$age^2),
##     data = doctor_general_practice)
##
## Residuals:
##       Min     1Q Median     3Q    Max 
## -4.1048 -0.1928  0.0874  0.2517  3.5931 
##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)               6.290e+00  3.500e-01 17.971
## doctor_general_practice$Scholarity -9.040e-02  3.623e-02 -2.495
## doctor_general_practice$age        -3.605e-03  6.251e-03 -0.577
## doctor_general_practice$gender    1.383e-02  1.924e-02  0.718
## doctor_general_practice$employment_time 1.018e-03  1.036e-04  9.833
## I(doctor_general_practice$age^2)    5.136e-06  6.822e-05  0.075
## 
## Pr(>|t|) 
## (Intercept) <2e-16 ***
## doctor_general_practice$Scholarity 0.0126 * 
## doctor_general_practice$age        0.5642  
## doctor_general_practice$gender    0.4725  
## doctor_general_practice$employment_time <2e-16 ***
## I(doctor_general_practice$age^2)    0.9400  
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.572 on 4208 degrees of freedom
## Multiple R-squared:  0.02788, Adjusted R-squared:  0.02672
## F-statistic: 24.13 on 5 and 4208 DF,  p-value: < 2.2e-16

```

Economist

```
reg = lm(data = economist, formula = log(economist$norm_avg_salary) ~ +
          economist$Scholarity + economist$age + economist$gender +
          economist$employment_time + I(economist$age^2))
summary(reg)

##
## Call:
## lm(formula = log(economist$norm_avg_salary) ~ +economist$Scholarity +
##      economist$age + economist$gender + economist$employment_time +
##      I(economist$age^2), data = economist)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -2.33181 -0.37372 -0.00991  0.34861  2.30300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.1982467  0.6693890  0.296  0.76717  
## economist$Scholarity 0.3219963  0.0724498  4.444 9.85e-06 ***
## economist$age      0.0597791  0.0130188  4.592 4.98e-06 *** 
## economist$gender     -0.2519291  0.0383869 -6.563 8.64e-11 *** 
## economist$employment_time 0.0029719  0.0002283 13.019 < 2e-16 ***
## I(economist$age^2)   -0.0005408  0.0001648 -3.281  0.00107 ** 
## ---                
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5627 on 955 degrees of freedom
## Multiple R-squared:  0.468, Adjusted R-squared:  0.4652 
## F-statistic: 168 on 5 and 955 DF, p-value: < 2.2e-16
```

Lawyer

```
reg = lm(data = lawyer, formula = log(lawyer$norm_avg_salary) ~ +
          lawyer$Scholarity + lawyer$age + lawyer$gender +
          lawyer$employment_time + I(lawyer$age^2))
summary(reg)

##
## Call:
## lm(formula = log(lawyer$norm_avg_salary) ~ +lawyer$Scholarity +
##      lawyer$age + lawyer$gender + lawyer$employment_time + I(lawyer$age^2),
##      data = lawyer)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -2.58409 -0.45657  0.01525  0.44757  3.10190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.100e+00  6.762e-01  1.627  0.1038  
## lawyer$Scholarity 1.686e-01  7.374e-02  2.286  0.0223 * 
## ---
```

```

## lawyer$age           8.612e-02  8.179e-03  10.530 < 2e-16 ***
## lawyer$gender        -1.437e-01  2.740e-02  -5.243 1.71e-07 ***
## lawyer$employment_time 3.664e-03  1.691e-04  21.666 < 2e-16 ***
## I(lawyer$age^2)      -8.866e-04  9.448e-05  -9.383 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6572 on 2470 degrees of freedom
## Multiple R-squared:  0.3252, Adjusted R-squared:  0.3239
## F-statistic: 238.1 on 5 and 2470 DF,  p-value: < 2.2e-16

```

Street Cleaner

```

reg = lm(data = street_cleaner, formula = log(street_cleaner$norm_avg_salary) ~ +
          street_cleaner$Scholarity + street_cleaner$age + street_cleaner$gender +
          street_cleaner$employment_time + I(street_cleaner$age^2))
summary(reg)

##
## Call:
## lm(formula = log(street_cleaner$norm_avg_salary) ~ +street_cleaner$Scholarity +
##     street_cleaner$age + street_cleaner$gender + street_cleaner$employment_time +
##     I(street_cleaner$age^2), data = street_cleaner)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -1.8352 -0.1646 -0.0456  0.1209  3.3346 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                2.890e+00  1.608e-02 179.68 <2e-16 ***
## street_cleaner$Scholarity  2.222e-02  9.159e-04  24.26 <2e-16 ***
## street_cleaner$age          1.187e-02  7.656e-04  15.50 <2e-16 ***
## street_cleaner$gender       -2.272e-01  3.135e-03 -72.46 <2e-16 ***
## street_cleaner$employment_time 2.073e-03  2.241e-05  92.54 <2e-16 ***
## I(street_cleaner$age^2)     -1.525e-04  9.190e-06 -16.60 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2822 on 38506 degrees of freedom
## Multiple R-squared:  0.2881, Adjusted R-squared:  0.288 
## F-statistic: 3116 on 5 and 38506 DF,  p-value: < 2.2e-16

```

Discussion

It seems that Mincer model for the normal logarithm of the normalized salary was well fitted for most of the professions, the profession doctor the drastic case where it seems to only have significant relation the employment time and maybe scholarship.

In the plot below we can see the difference of the fitted model considering only age.

```

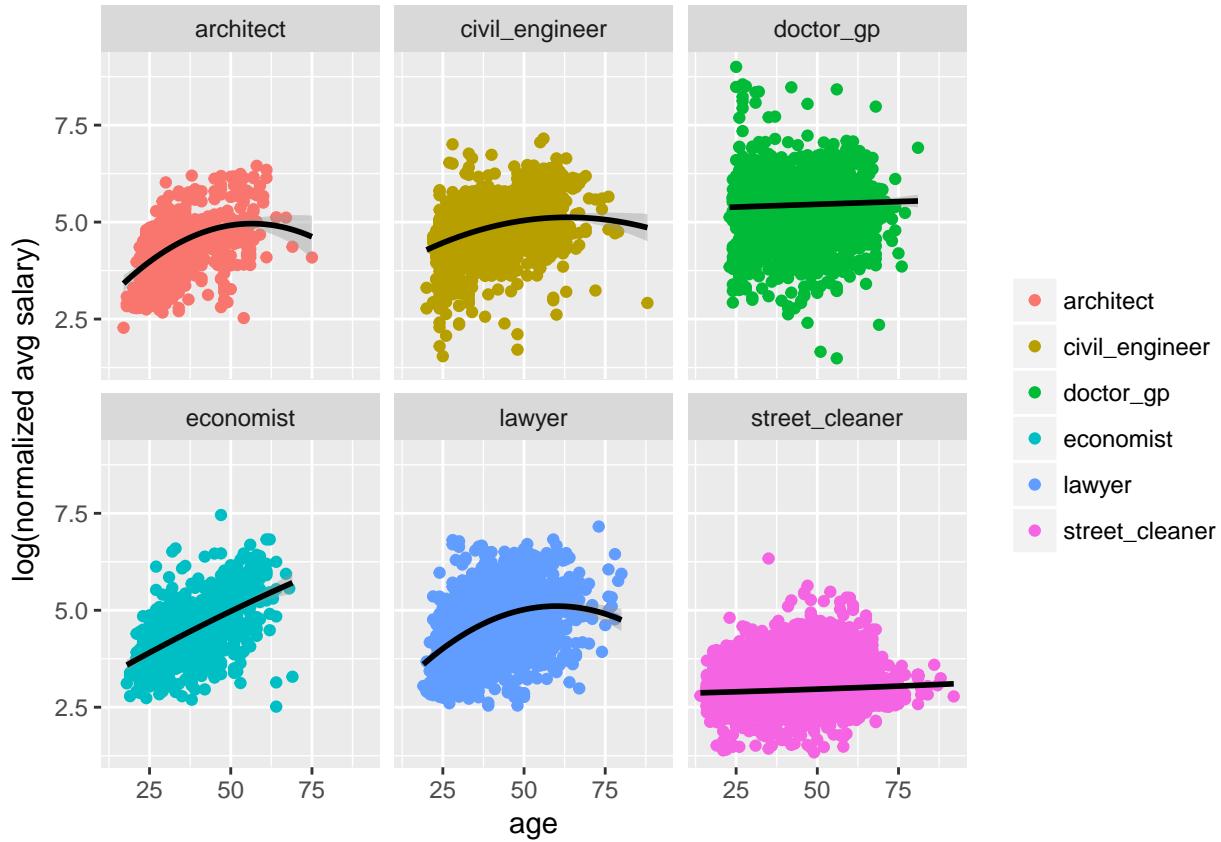
plot = ggplot(data = dataset, aes(dataset$age, log(dataset$norm_avg_salary),
                                    color = dataset$CBO2002)) + geom_point() +

```

```

geom_smooth(color = "black", method = lm, formula = y ~ x + I(x^2)) + labs(x = "age", y = "log(normalized avg salary")
facet_wrap(~ dataset$CB02002) + theme(legend.title=element_blank())
plot

```



Now, with these models, I will try to answer the questions that I presented in the beginning of the document:

1. Is there a difference between the average salary between genders? If it is the case, in which profession we have the biggest salary gap per gender?

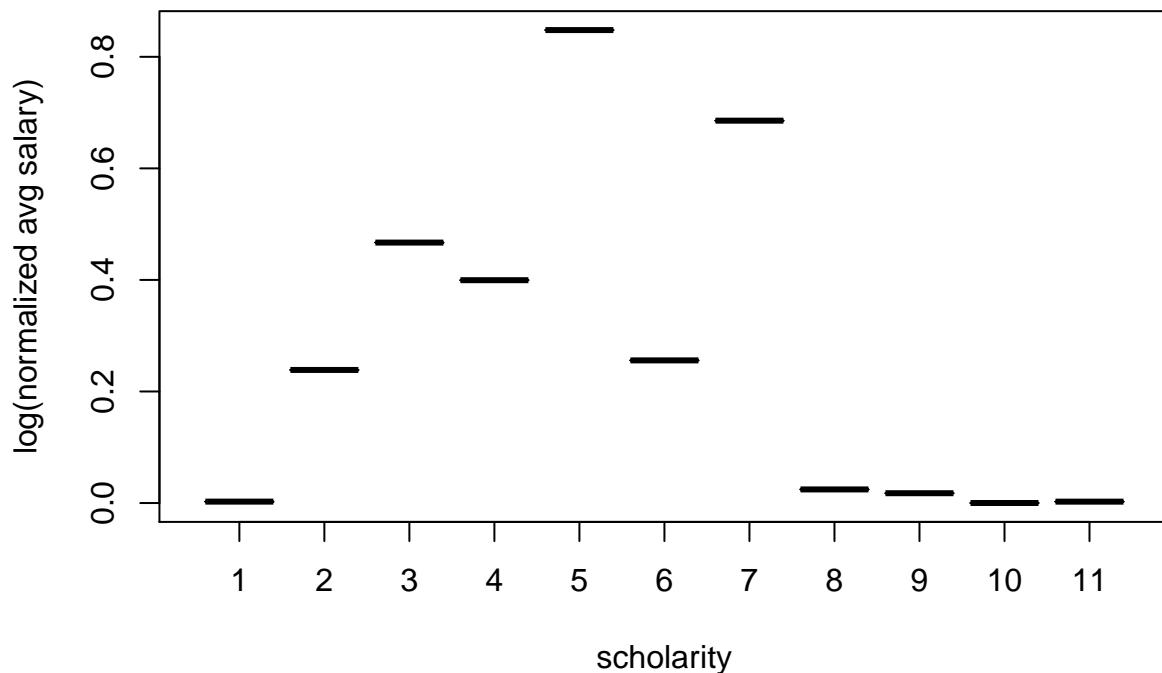
The regressions (excluding the doctor dataset) gave us the prove that the gender is a variable that helps us to explain the model, which means it has an impact in the salary. Then, with this information we can calculate the difference between the salary for a woman and a man in a profession, to find the biggest gap:

- architect: -0.0116690918024
- civil engineer: 0.139170680059
- **economist: 0.521903520552**
- lawyer: 0.288922295832
- street cleaner: 0.230664127093

These values are in percentage. The profession economist, where we have the biggest salary gap, is a field with discrepancy close to 0.52% between salaries. At the same time, it is interesting to notice the profession architect, where women gain more than men (approximately 0.01%).

2. What is the impact that scholarly have in the average salary? In our datasets, the street cleaner seems to be the only profession with representability for all levels of scholarly. There, we found a considerable increase between 4 to 5 (0.44%), and also a considerable increase between 6 and 7 (0.42%). Besides that, what we have as information is that this variable impacts positively the salary.

```
load(file="data/avg_street_cleaner.Rdata")
plot(factor(avg_street_cleaner$V1), avg_street_cleaner$V2, ylab = "log(normalized avg salary)", xlab = "scholarity")
```



3. How does the age affect the salary?

Conclusion

Because of the law structure in Brazil, it's possible to have underrepresentation for some professions (like doctors and lawyers, that sometimes register as partners in their business), and sometimes we also have a problem concerning the profession used for the registration, since sometimes a professional can be registered in two different ways (e.g. economists sometimes are registered as "analysts").

References

- [1] Sites about Mincer Earnings Function:

https://en.wikipedia.org/wiki/Mincer_earnings_function

<http://eml.berkeley.edu/~cle/wp/wp62.pdf>

[\(in portuguese\)](http://www.cps.fgv.br/cps/pesquisas/Politicas_sociais_alunos/2011/pdf/BES_EquacaoMinceriana.pdf)

- [2] Removing statistically significant intercept term increases R^2 in linear model:

<http://stats.stackexchange.com/questions/26176/removal-of-statistically-significant-intercept-term-increases-r2-in-linear-mo>