

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).

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:

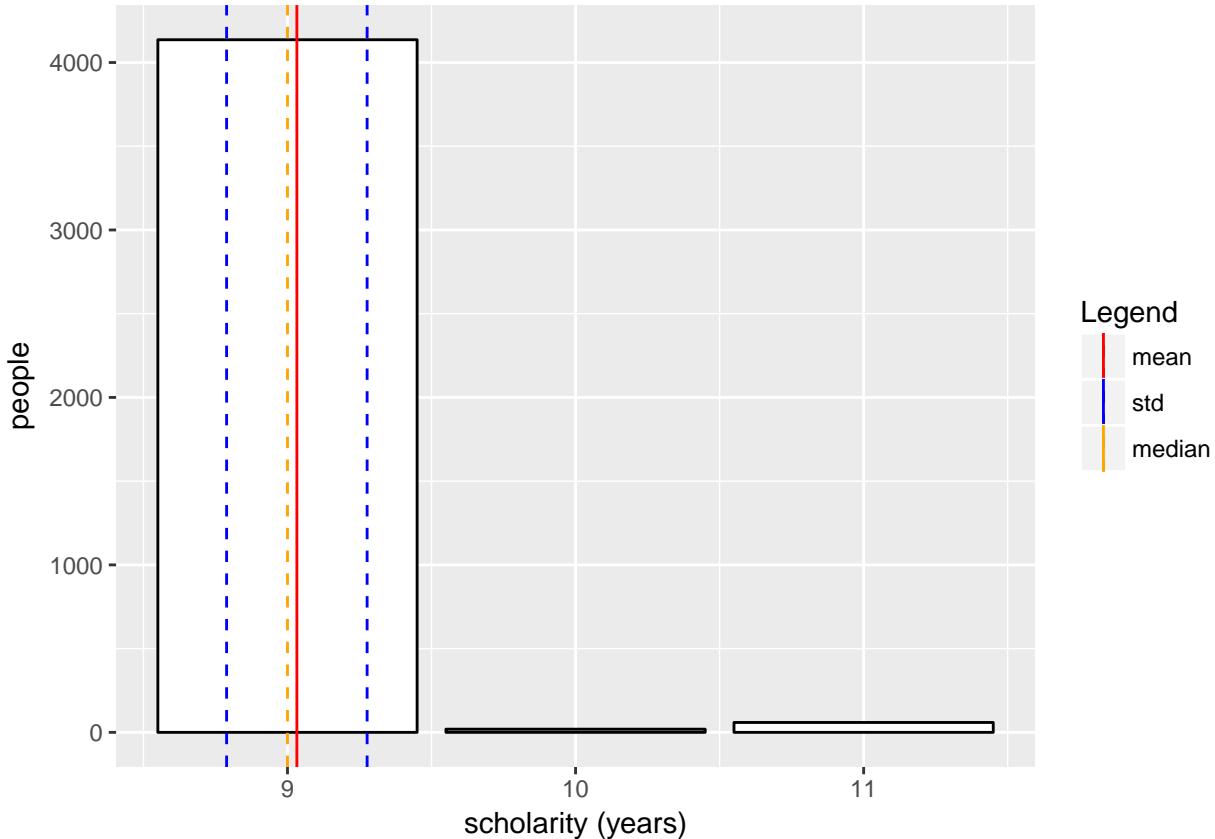
	Scholarity	Age	Contract hours	Employment time	Min salary	Avg salary
architect	Min.	5	17	5.00	0.60	272.50
	1 st Qu.	9	28	30.00	7.90	2,038.80
	Median	9	32	44.00	23.80	6.00
	Mean	9	35.76	37.91	59.51	7.30
	3 rd Qu.	9	43	44.00	58.85	8.97
	Max.	10	75	44.00	478.90	41.98
civil engineer	Min.	3	20	3	0.30	168.20
	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")
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
```



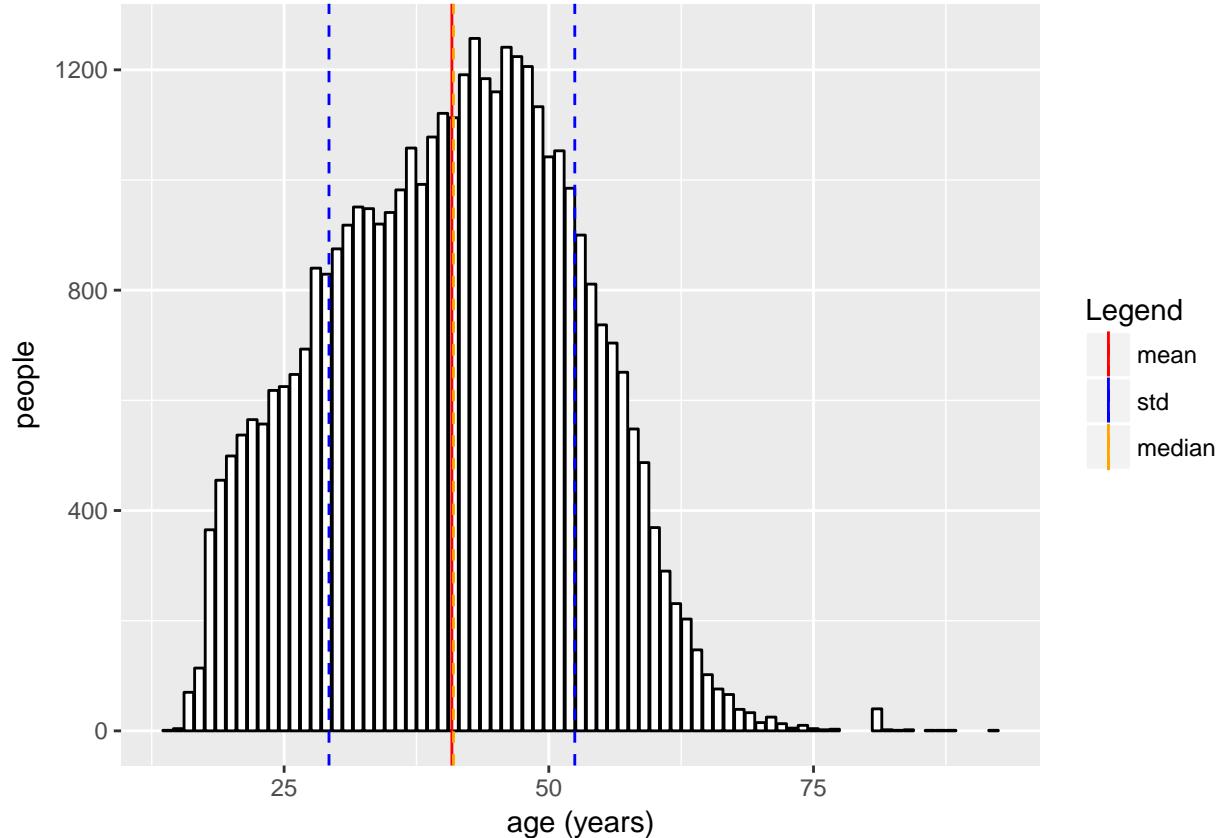
- Age:

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 profesisons, 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
```

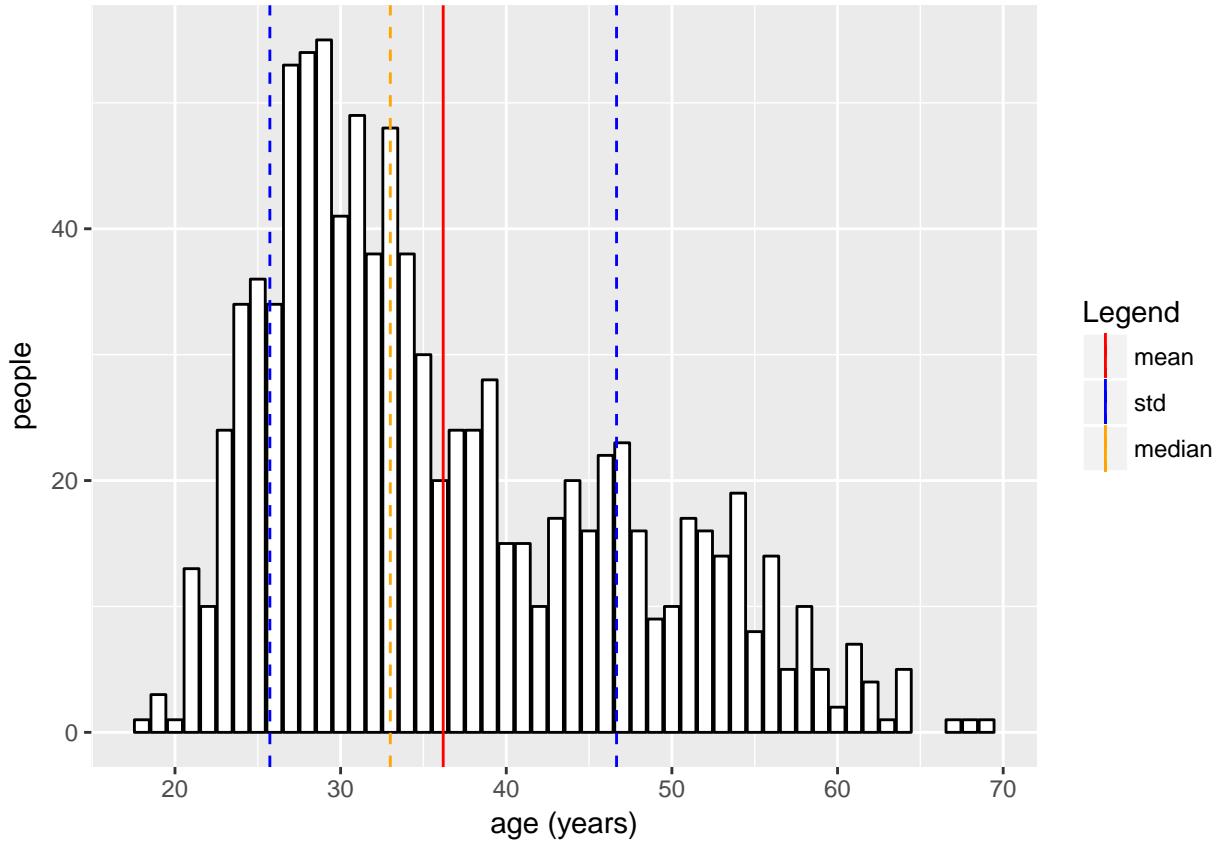


```
meanE <- mean(economist$age)
std <- sd(economist$age)
economistPlot = ggplot(data = economist, aes(economist$age)) +
  geom_bar(fill="white", colour = "black") + labs(x= "age (years)", y = "people")+
  geom_vline(aes(xintercept = meanE, colour = "mean")) +
  geom_vline(aes(xintercept = (meanE + std),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

- **Employment Time:**

Employment time represents the quantity of months that the employee worked registered in the profession in question. 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.

```

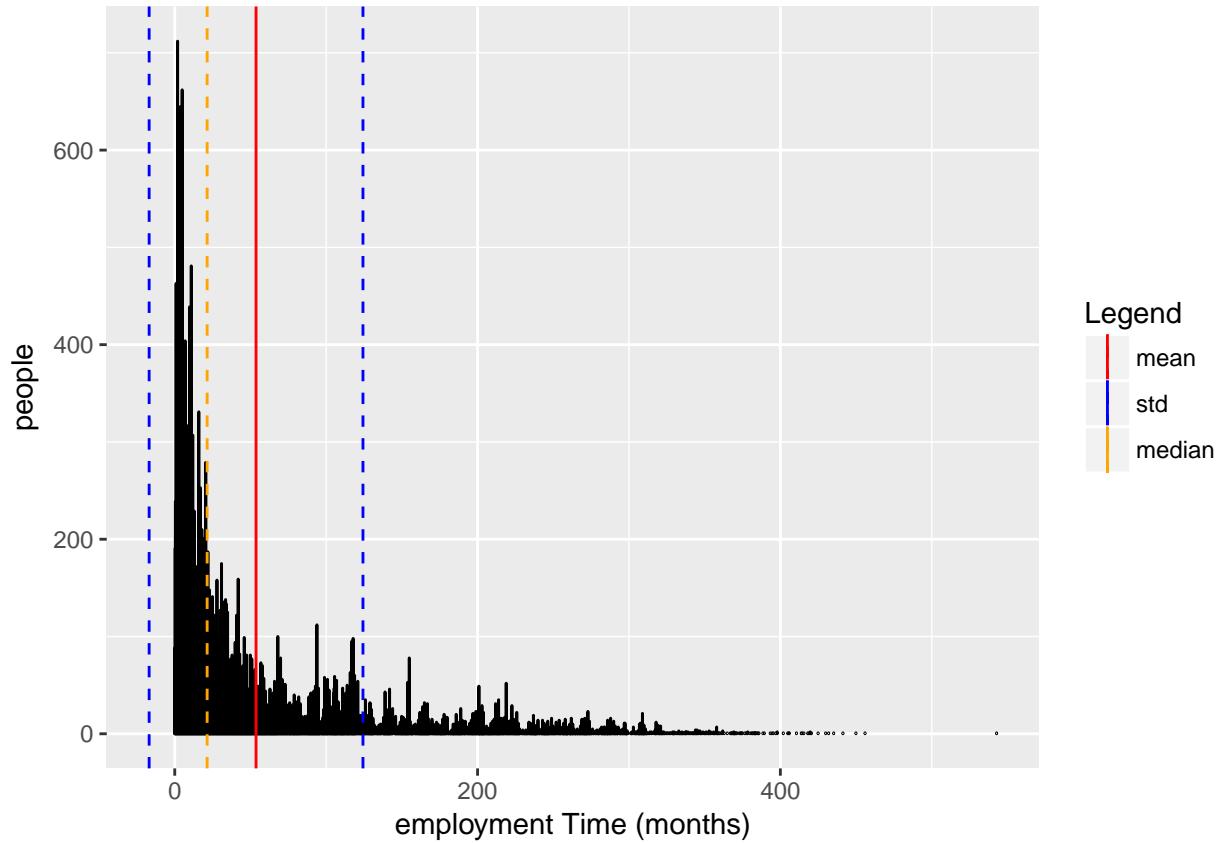
meanE <- mean(street_cleaner$employment_time)
std <- sd(street_cleaner$employment_time)
streetPlot = ggplot(data = street_cleaner, aes(street_cleaner$employment_time)) +
  geom_bar(fill="white", colour = "black") + labs(x= "employment Time (months)", y = "people") +
  geom_vline(aes(xintercept = meanE,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

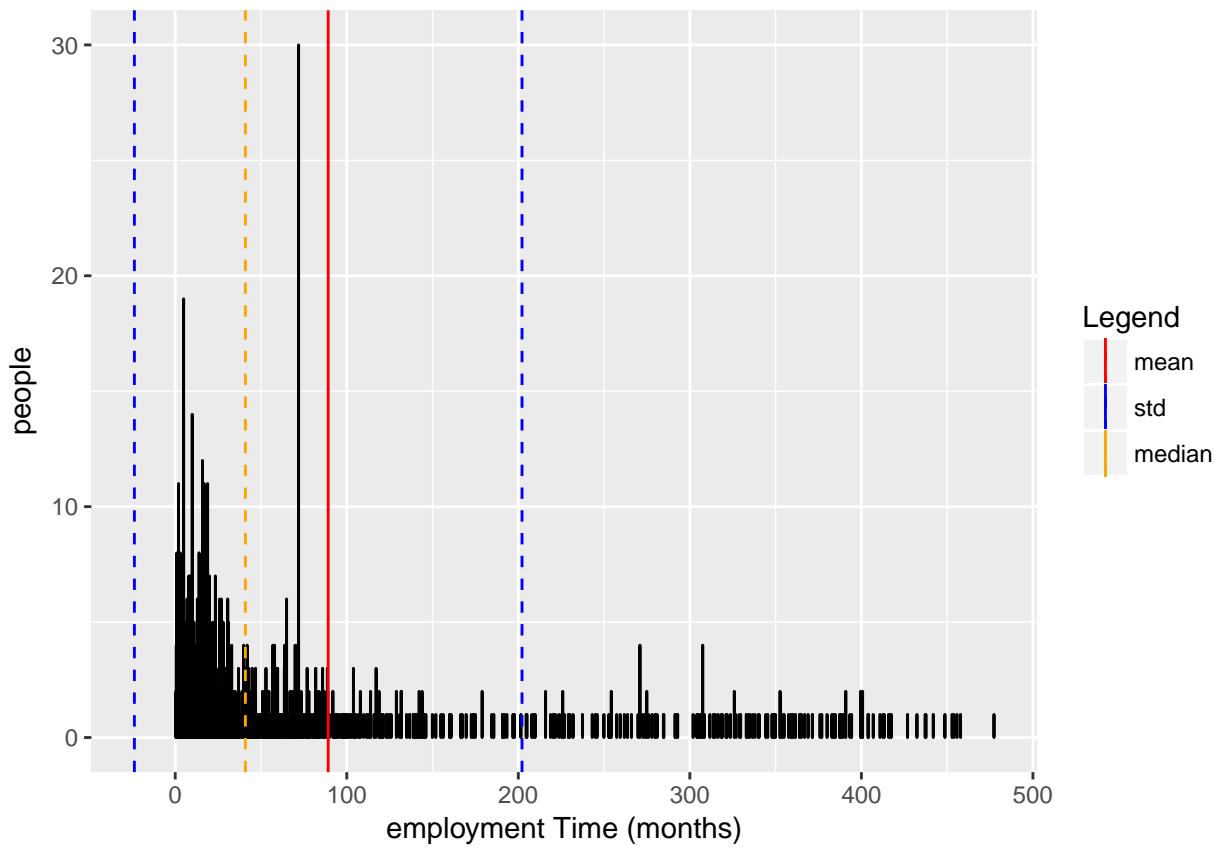
```



```

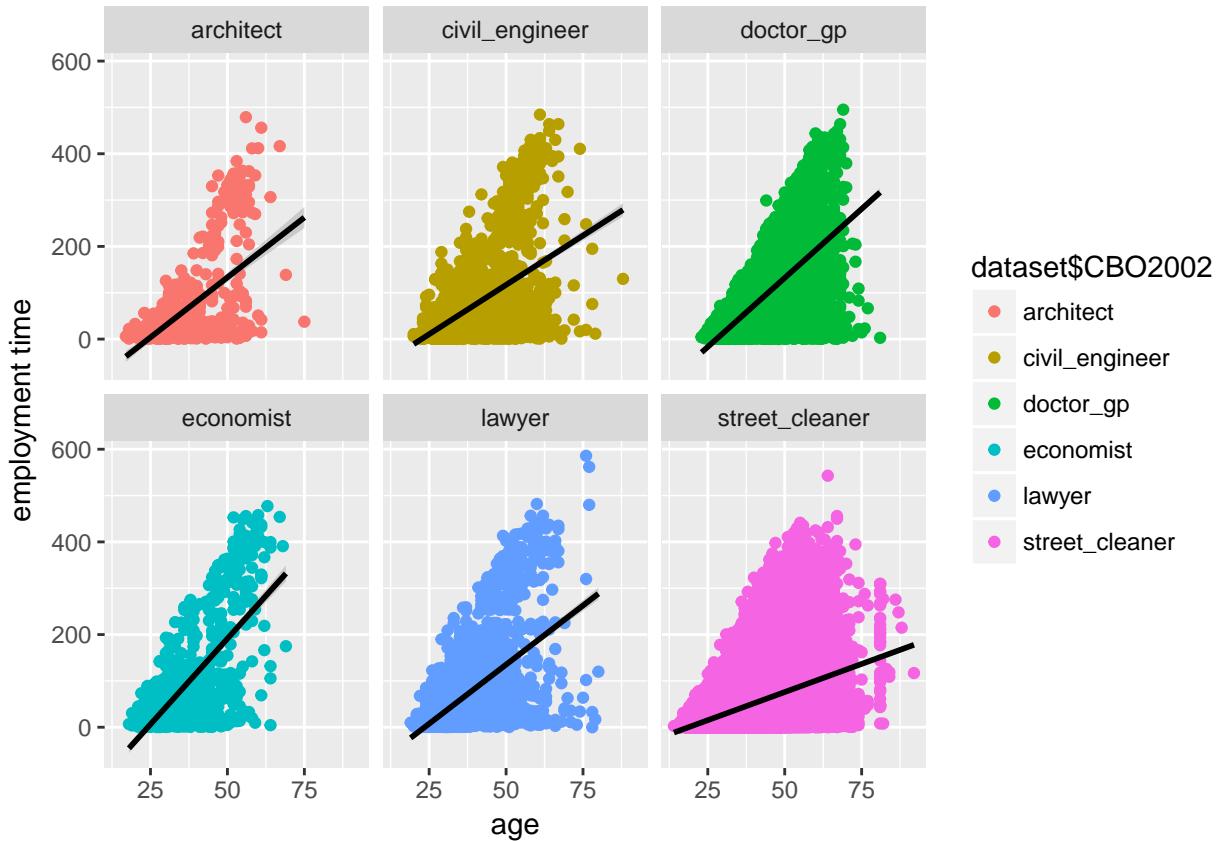
meanE <- mean(economist$employment_time)
std <- sd(economist$employment_time)
economistPlot = ggplot(data = economist, aes(economist$employment_time)) +
  geom_bar(fill="white", colour = "black") + labs(x= "employment Time (months)", y = "people")+
  geom_vline(aes(xintercept = meanE, colour = "mean")) +
  geom_vline(aes(xintercept = (meanE + std), colour = "std"), linetype = "dashed") +
  geom_vline(aes(xintercept = (meanE - std), colour = "std"), linetype = "dashed") +
  geom_vline(aes(xintercept = median(economist$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$CB02002)) +
  geom_point() + labs(y = "employment time", x = "age") + geom_smooth(method= "lm",color= "black")+
  facet_wrap(~ dataset$CB02002)
plot
```



It looks difficult to fit a model with so much variability for each age. Below, it's possible to see the summary for the regression (terrible adjusted R-squared value!).

```
reg = lm(dataset$employment_time~dataset$age)
summary(reg)
```

```
##
## Call:
## lm(formula = dataset$employment_time ~ dataset$age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -177.79  -44.49  -12.62   21.53  420.43 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -62.87462   1.16764 -53.85 <2e-16 ***
## dataset$age    3.00456   0.02767 108.57 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 71.19 on 48999 degrees of freedom
## Multiple R-squared:  0.1939, Adjusted R-squared:  0.1939 
## F-statistic: 1.179e+04 on 1 and 48999 DF,  p-value: < 2.2e-16
```

Eu não sei se fez sentido o que eu fiz aqui...

- **Min Salary:**

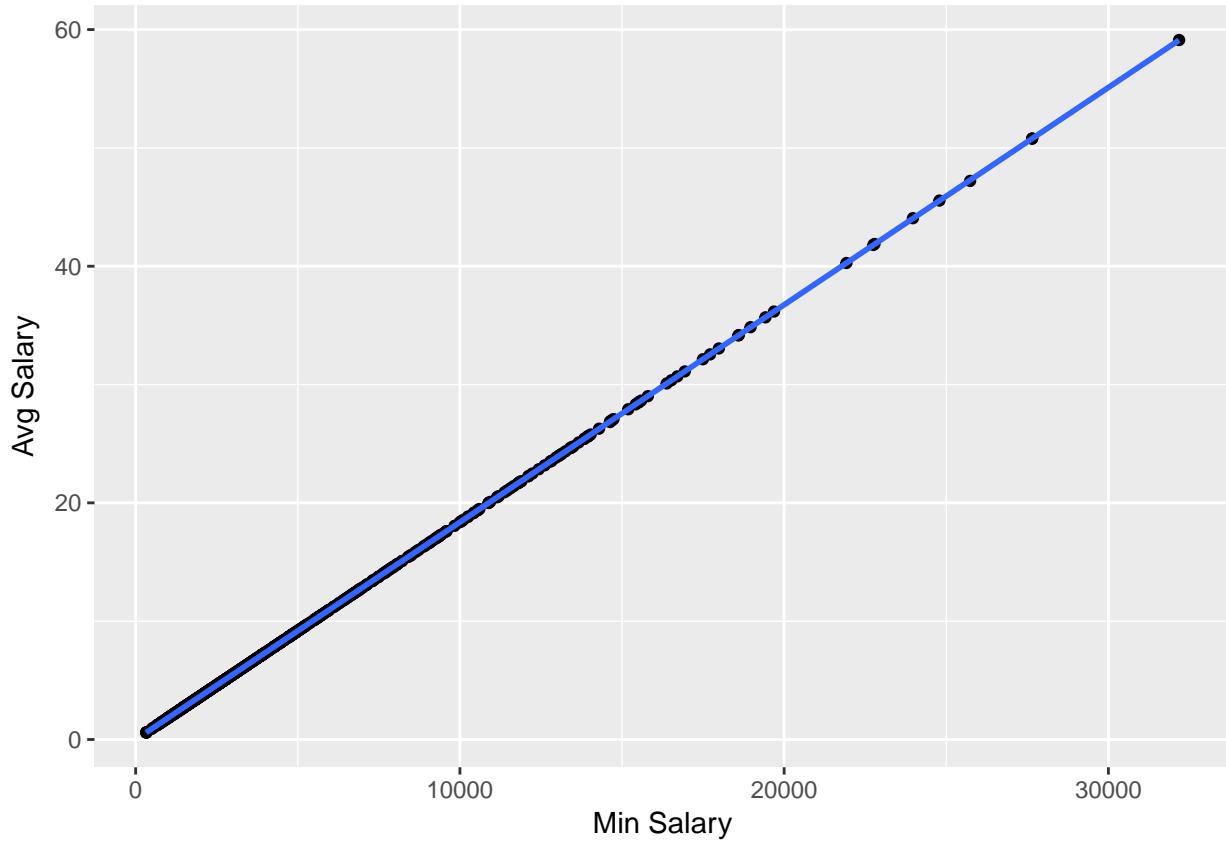
This variable represent the relation between the employee salary and the brazilian minimum wage. In 2014,

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.

Variables of interest against Average Salary

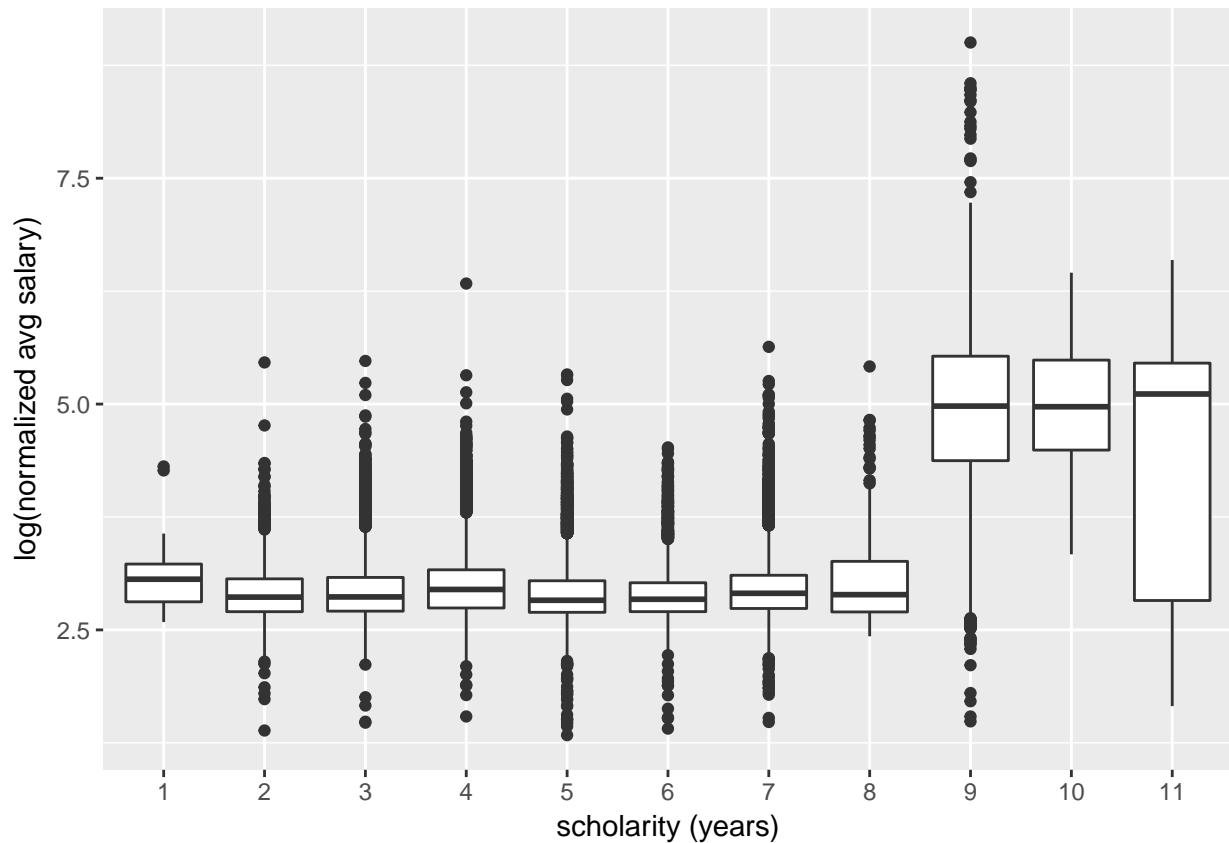
Before jumping to comparisons between salary and other variables, I normalize the average salary to allow a fair comparison between the observation in our datasets. This normalization is needed because without it we could conclude that some cases in our dataset are receiving more or less money because 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 an range transformation. In order to better visualize and summarize the normalized average salary, for this analysis we are going to apply a logarithm on it, since it does not change the values order, just scales it.

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
```

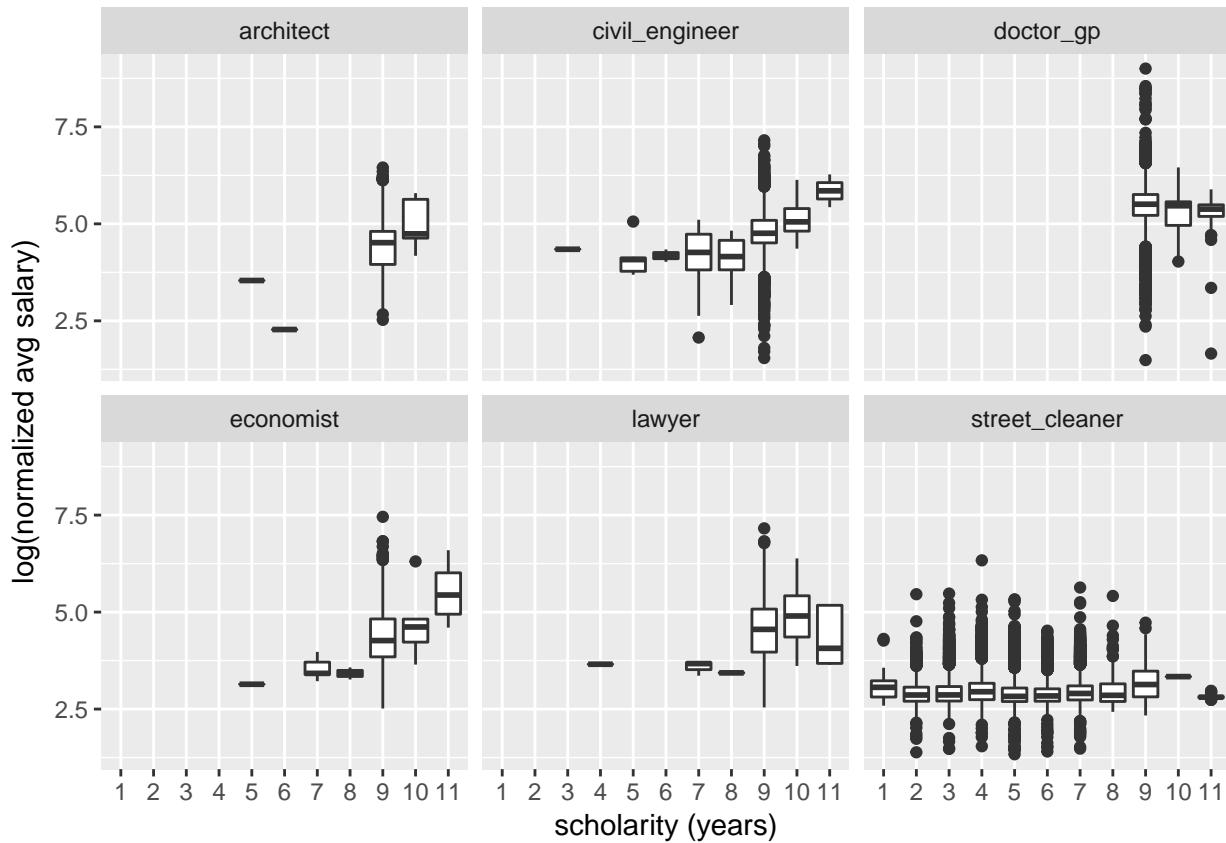


This visualization help us to identify a possible relationship between the salary and the scholarly. Although between the first seven years of scholarly we do not seem to have a considerable difference, from the eighth to the ninth year, it seems to impact drastically the average salary of the employee.

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
```



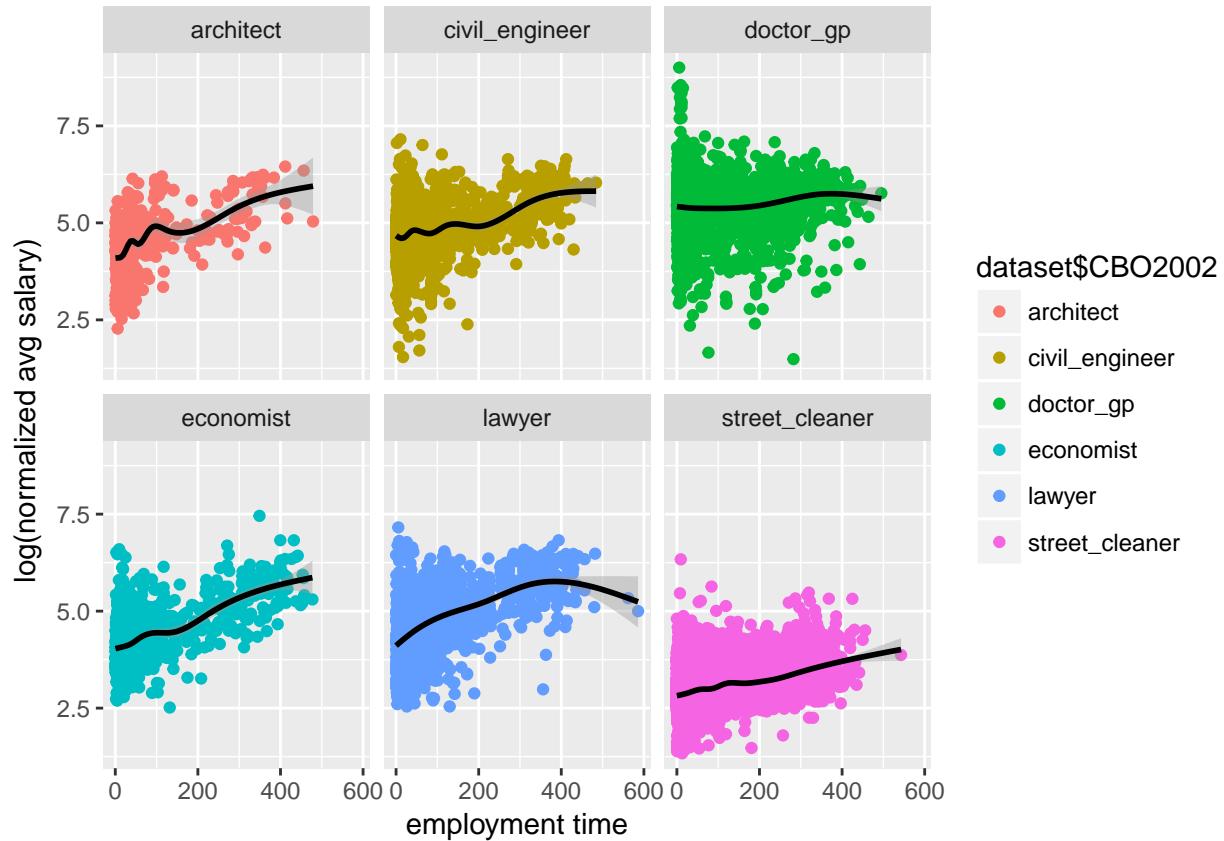
Age

```
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)
```

Employment Time

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

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



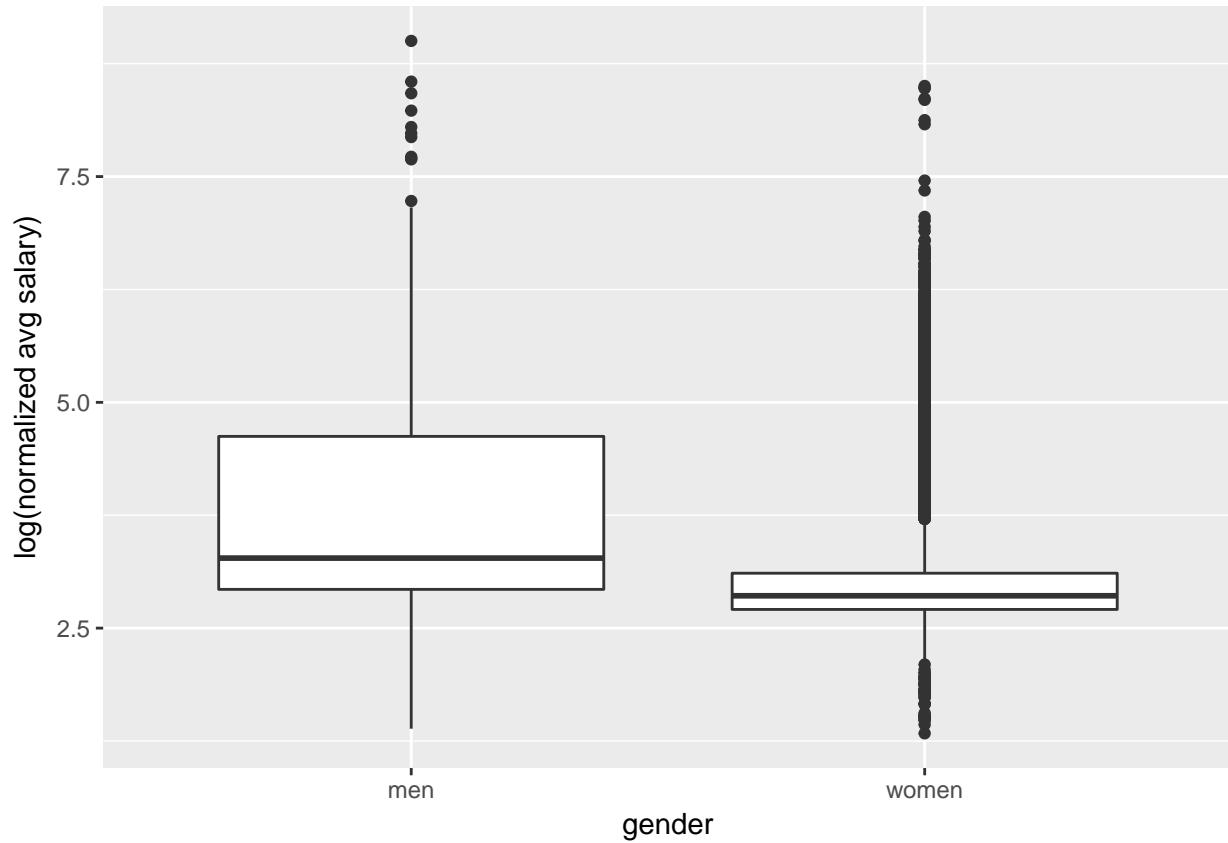
Similar as when comparing age with employment time, the observations seem to have a huge variability.

como eu faço pra lidar com essa nuvem gigante de pontos? DDDDD=

Se tu souberes como mudar o nome dessa legenda, eu vou ficar bem feliz, pq fiquei uma hora tentando

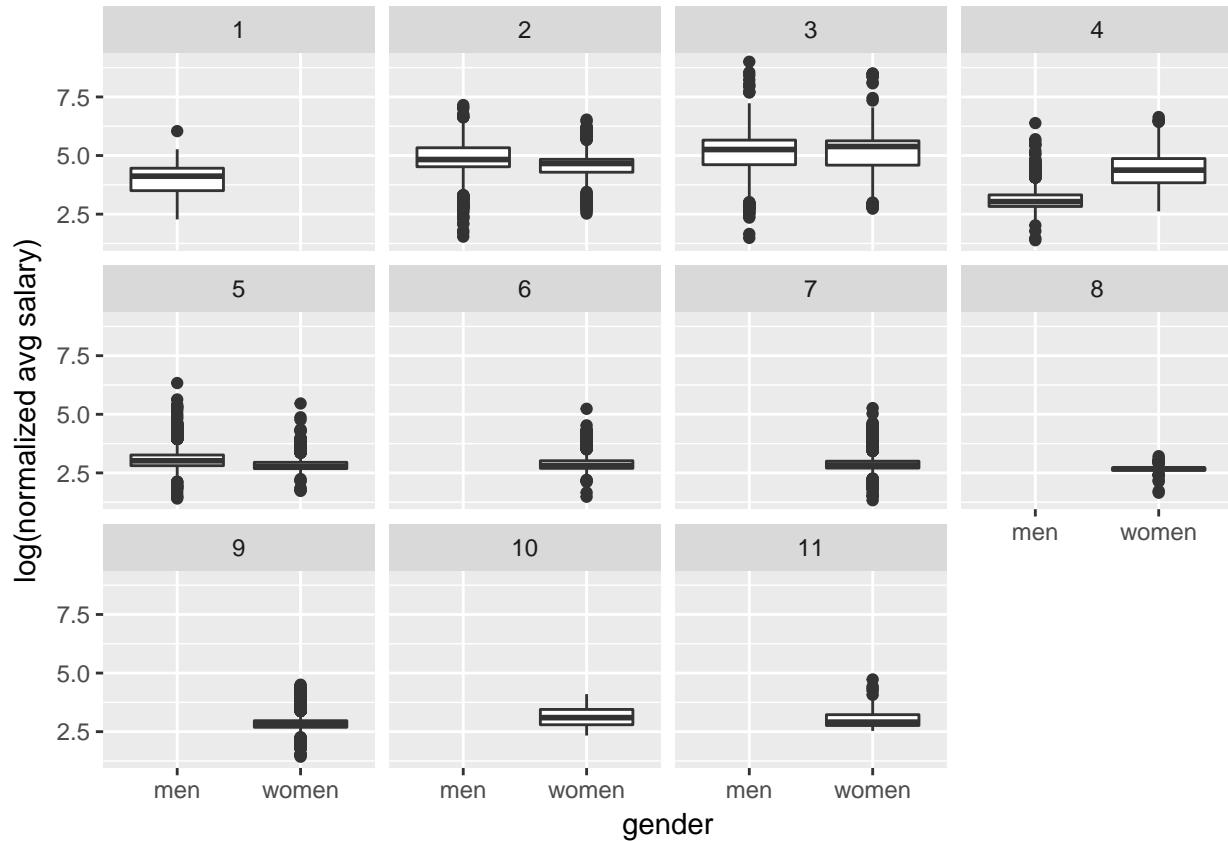
Gender

```
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
```



não me diz nada, tenho q plotar por escolaridade e profissão!

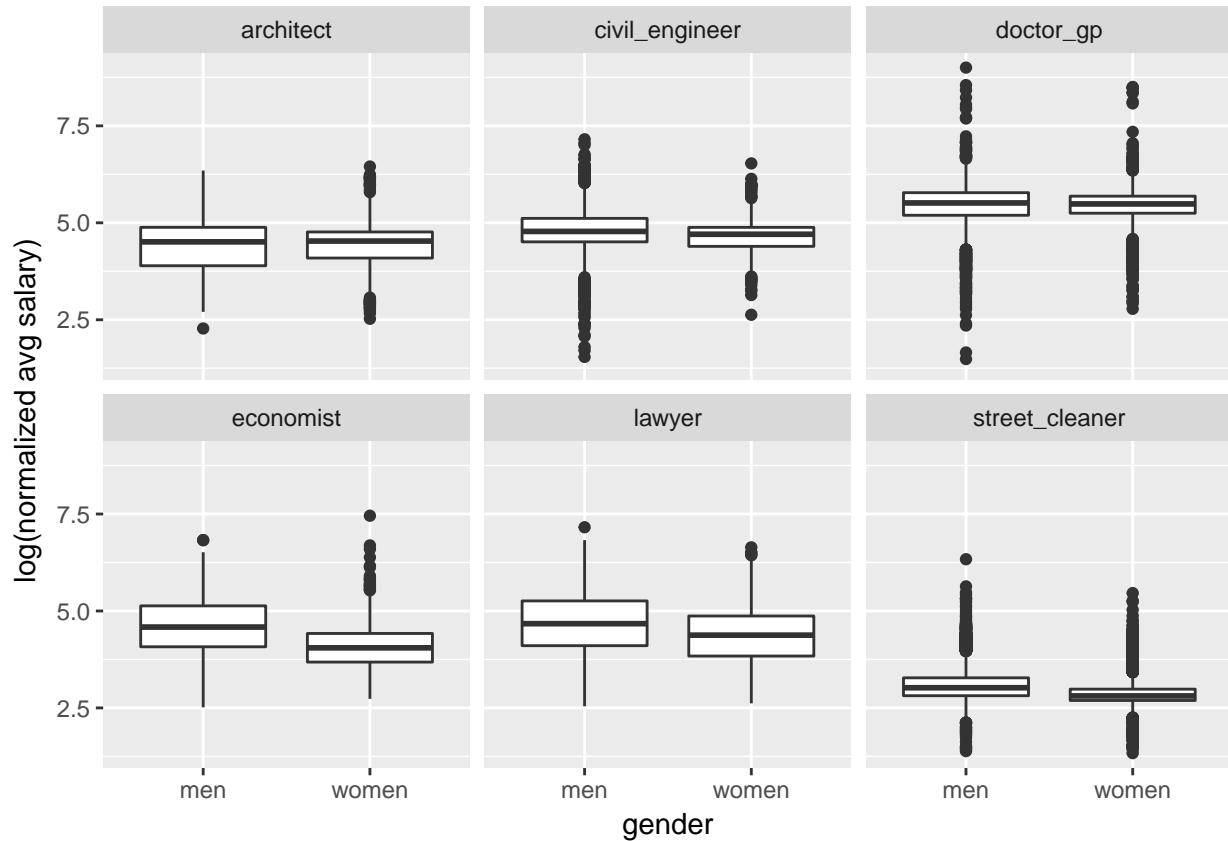
```
plot = ggplot(data=dataset, aes(x= factor(dataset$gender, labels=c("men", "women")),
                                 y=log(dataset$norm_avg_salary))) +
  geom_boxplot() + labs(x = "gender", y = "log(normalized avg salary)") +
  facet_wrap(~ dataset$Scholarity)
plot
```



NÃO TEM HOMEM COM ESCOLARIDADE MAIOR QUE 6?????????????????????????????

Isso não faz sentido! Médico é só com escolaridade acima de 9, e no plot de baixo dá pra ver que tem um monte de homem médico!

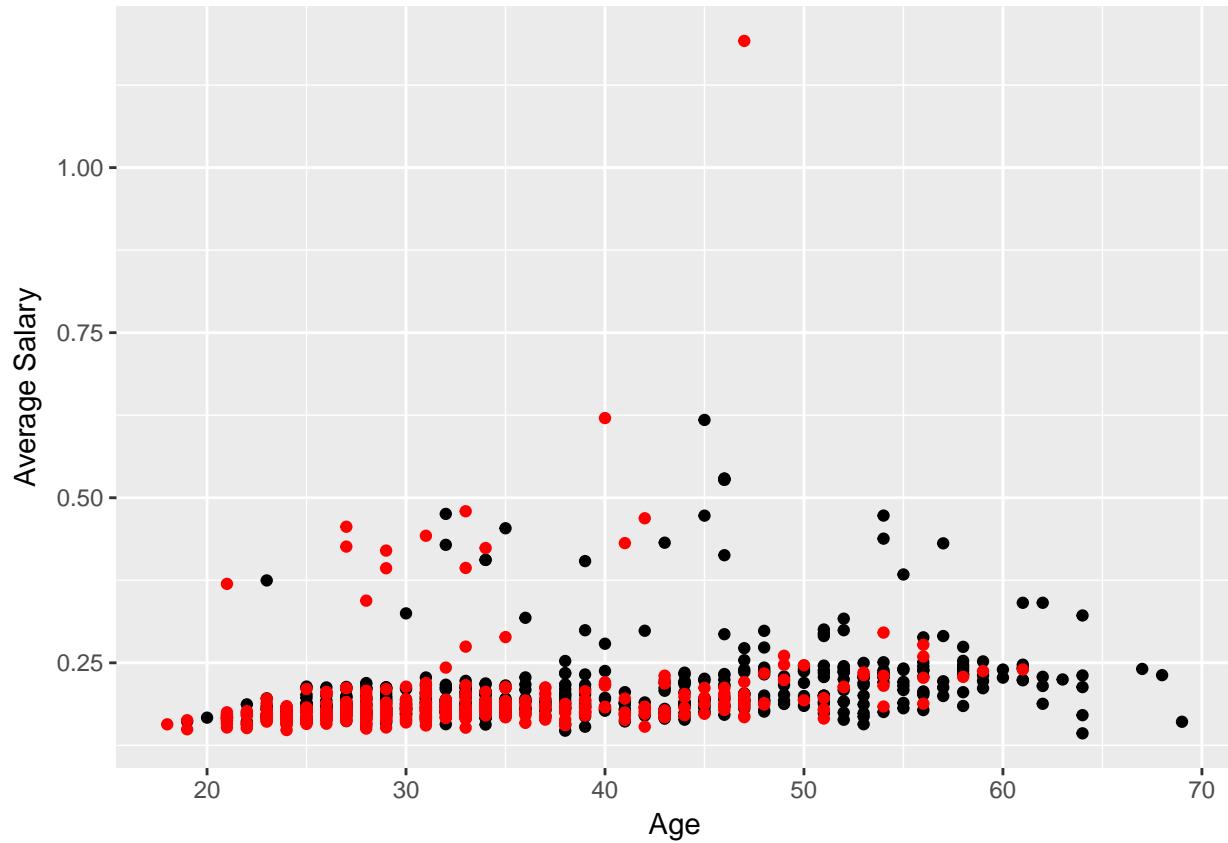
```
plot = ggplot(data=dataset, aes(x= factor(dataset$gender, labels=c("men","women")),
                                 y=log(dataset$norm_avg_salary))) +
  geom_boxplot() + labs(x = "gender", y = "log(normalized avg salary)") +
  facet_wrap(~ dataset$CB02002)
plot
```



A PARTIR DAQUI É ANTIGO, NEM PERCA SEU TEMPO LENDO

Gender and Age against Average Salary

```
ggplot(data = economist, aes(economist$age, log(economist$avg_salary)/ economist$contract_hours )) + geom
```



Discussion

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").