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 specificaly 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, scholarity, profession, etc) can impact the average salary. More specificaly, 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 scholarity have in the average salary?
3. How does the age affect the salary?

Descritives

We have seven variables in each dataset: scholarity (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 bellow 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:

- **Scholarity:**

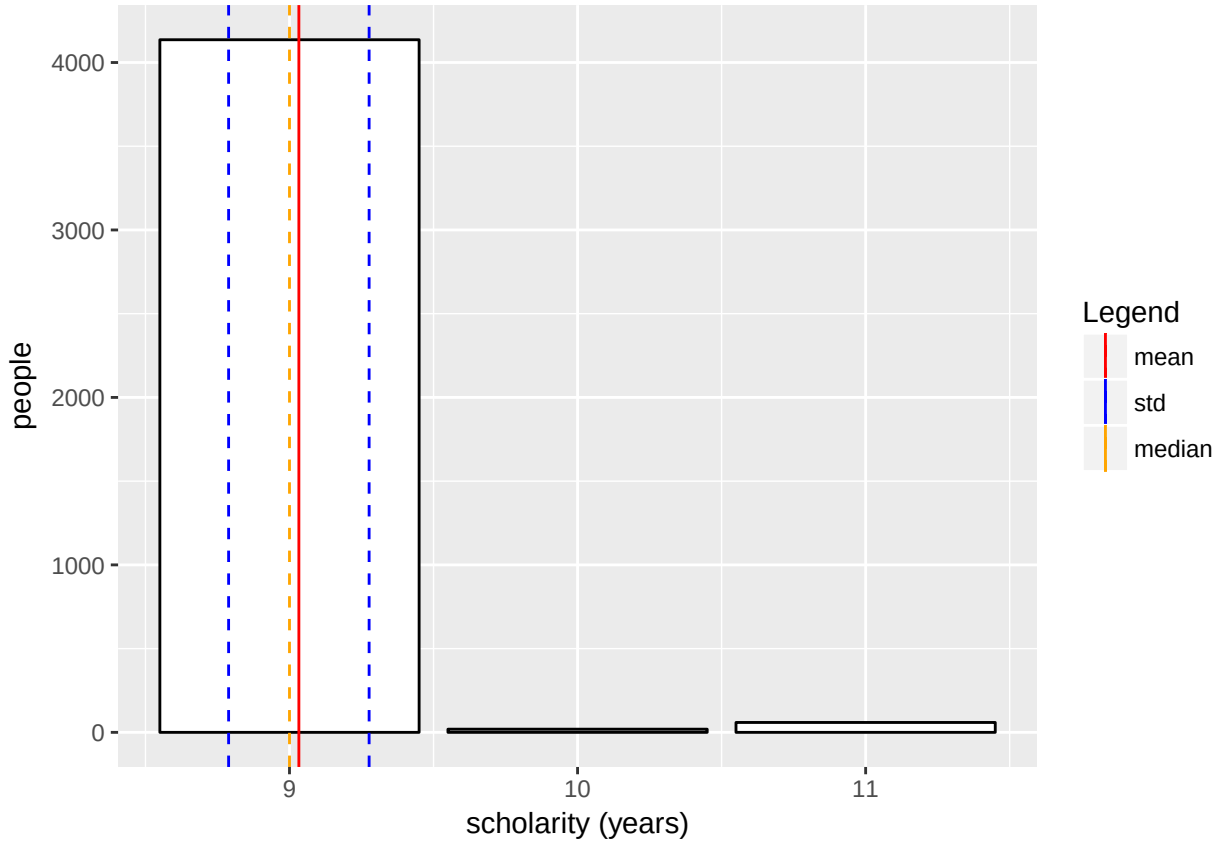
		Scholarity	Age	Contract hours	Employment time	Min salary	Avg salary
architect	Min.	5	17	5.00	0.60	0.50	272.50
	1st Qu.	9	28	30.00	7.90	3.73	2,038.80
	Median	9	32	44.00	23.80	6.00	3,270.00
	Mean	9	35.76	37.91	59.51	7.30	3,979.10
	3rd Qu.	9	43	44.00	58.85	8.97	4,888.00
	Max.	10	75	44.00	478.90	41.98	22,845.00
civil engineer	Min.	3	20	3	0.30	0.30	168.20
	1st Qu.	9	29	35	10.50	6.00	3,270.00
	Median	9	34	40	27.90	8.41	4,578.60
	Mean	8.939	37.72	38.07	64.65	9.70	5,285.80
	3rd Qu.	9	46	44	78.60	11.18	6,094.20
	Max.	11	88	44	484.40	93.93	51,194.40
doctor GP	Min.	9	23	1	0.20	0.32	177.00
	1st Qu.	9	33	20	10.20	7.04	3,838.00
	Median	9	41	22	39.40	11.40	6,213.00
	Mean	9.033	43.12	27.26	91.09	12.09	6,587.00
	3rd Qu.	9	54	40	134.80	16.81	9,160.00
	Max.	11	81	44	495.10	55.72	30,312.00
economist	Min.	5	18	8	0.40	0.58	323.70
	1st Qu.	9	28	40	15.70	3.66	2,000.00
	Median	9	33	44	40.90	5.60	3,057.00
	Mean	9	36.19	41.75	89.17	8.11	4,418.70
	3rd Qu.	9	44	44	101.90	9.28	5,061.80
	Max.	11	69	44	477.30	59.12	32,180.30
lawyer	Min.	4	19	1	0.00	0.30	167.90
	1st Qu.	9	30	40	11.90	3.64	1,987.30
	Median	9	34	40	33.90	6.31	3,440.80
	Mean	9	37.17	38.13	69.72	8.54	4,650.20
	3rd Qu.	9	43	44	78.33	10.53	5,738.90
	Max.	11	80	44	585.90	70.71	38,541.40
street cleaner	Min.	1	14	1	0.00	0.30	165.00
	1st Qu.	4	32	40	6.90	1.11	608.00
	Median	5	41	44	21.40	1.35	737.50
	Mean	4.904	40.85	40.63	53.70	1.50	818.90
	3rd Qu.	6	49	44	71.90	1.65	905.50
	Max.	11	92	44	542.80	15.81	8,610.90

Figure 1: Descritives 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
```



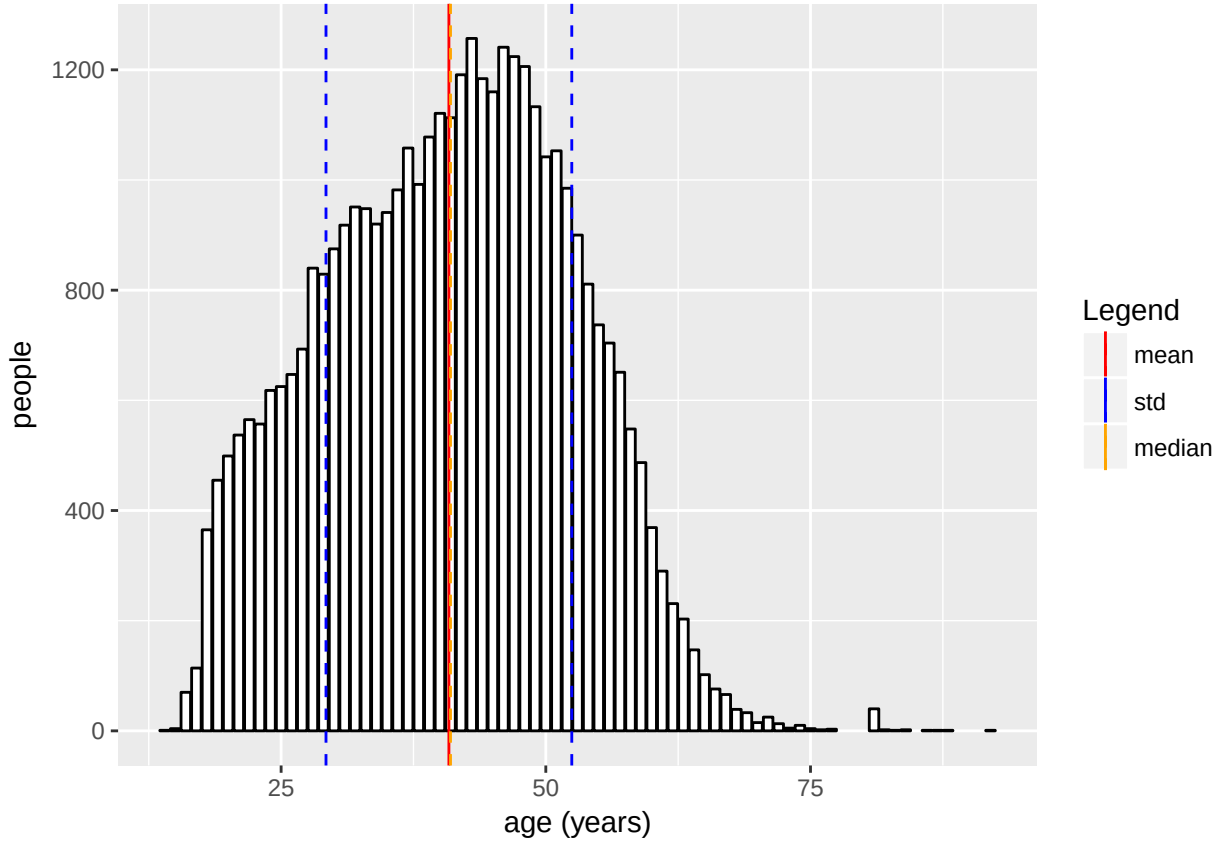
- **Age:**

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.

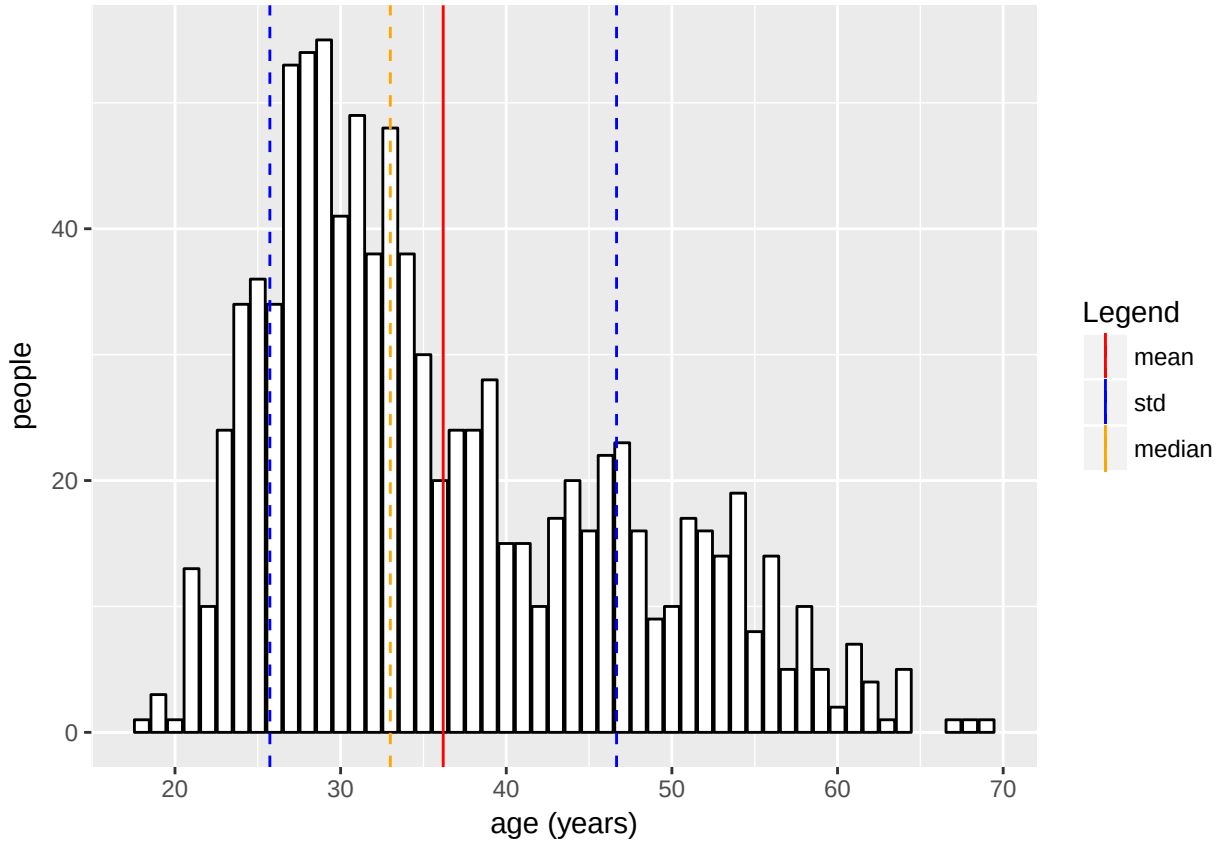
In Brazil, the age for retirement in 2014 was 59 years.

```
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 concentraring close to this number.

- **Gender:**

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

- **Employment Time:**

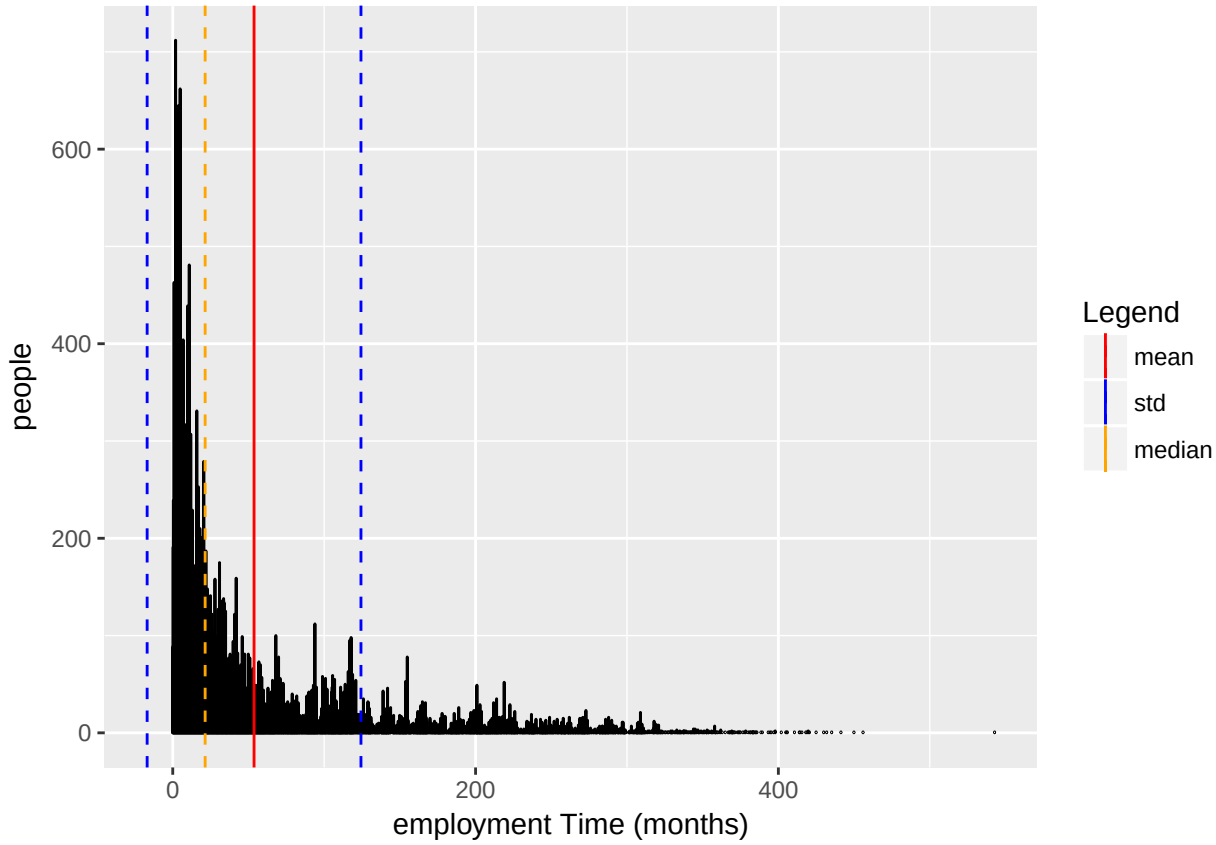
Employment time represents the quantity of months that the employee worked registered in a company. The graphs plotted below show the employment time per amount of people. The graphs show us that we have a greater amount of people with small employment time. Considering that in general we are supposed 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") +
  geom_vline(aes(xintercept = meanE, colour = "mean")) +
  geom_vline(aes(xintercept = (meanE + std), colour = "std"), linetype = "dashed") +
```

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%

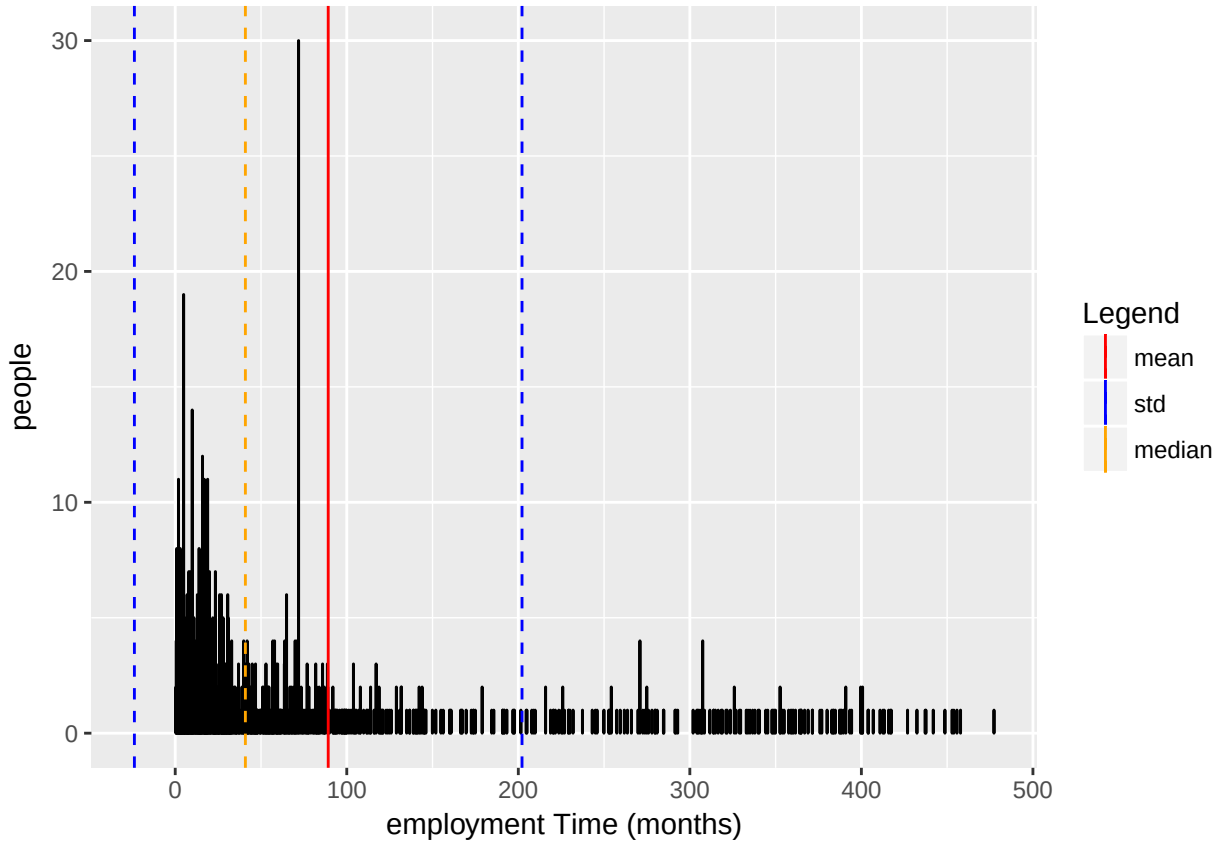
Figure 2: Gender percentage per profession

```
  geom_vline(aes(xintercept = (meanE - 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
```



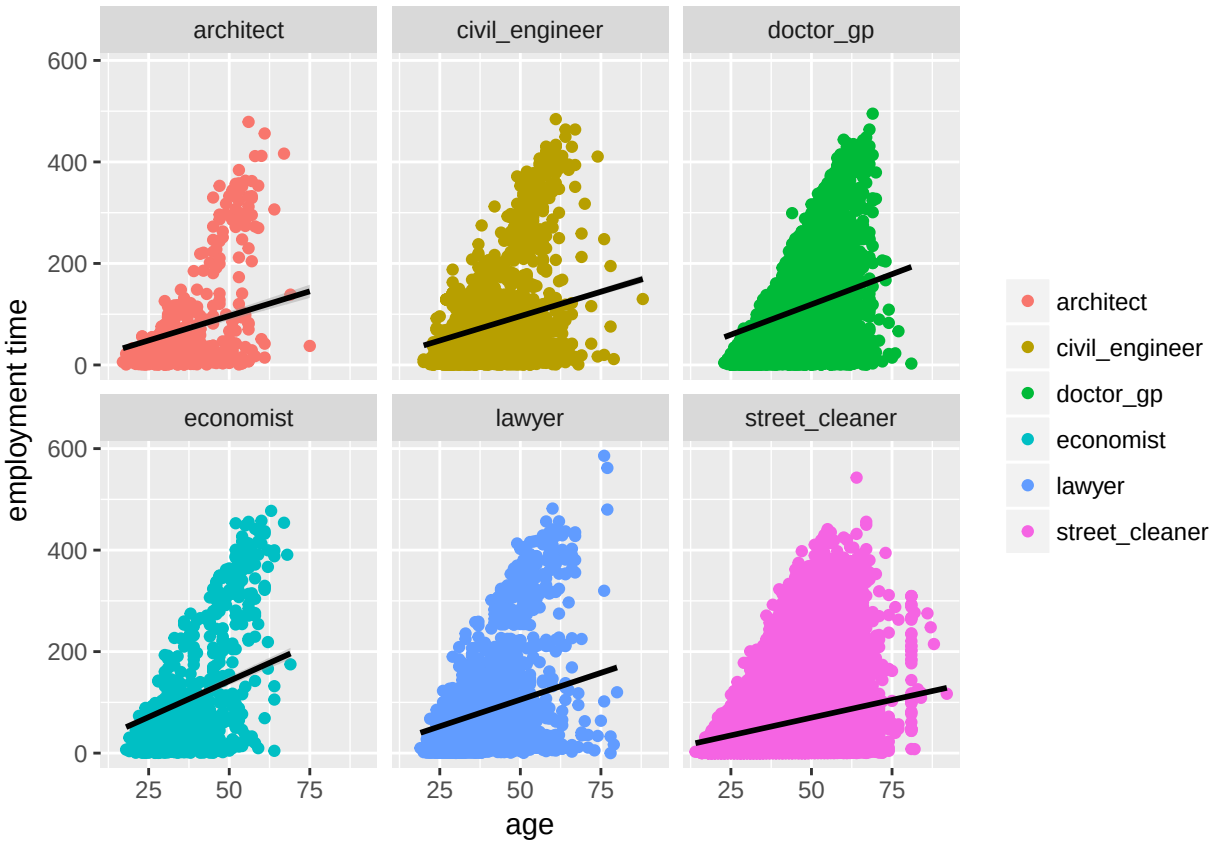

```
meanE <- mean(economist$employment_time)
std <- sd(economist$employment_time)
economistPlot = ggplot(data = economist, aes(economist$employment_time)) +
  geom_bar(fill="white", colour = "black") + labs(x= "employment Time (months)", y = "people")+
```

```
  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 ploted 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$CBO2002)) +
  geom_point() + labs(y = "employment time", x = "age") + geom_smooth(method= "lm",color=  "black", for
  facet_wrap(~ dataset$CBO2002) + theme(legend.title=element_blank())
plot
```



In this model, I removed the intecept, 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 bellow:

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

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

The variables are definately 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.

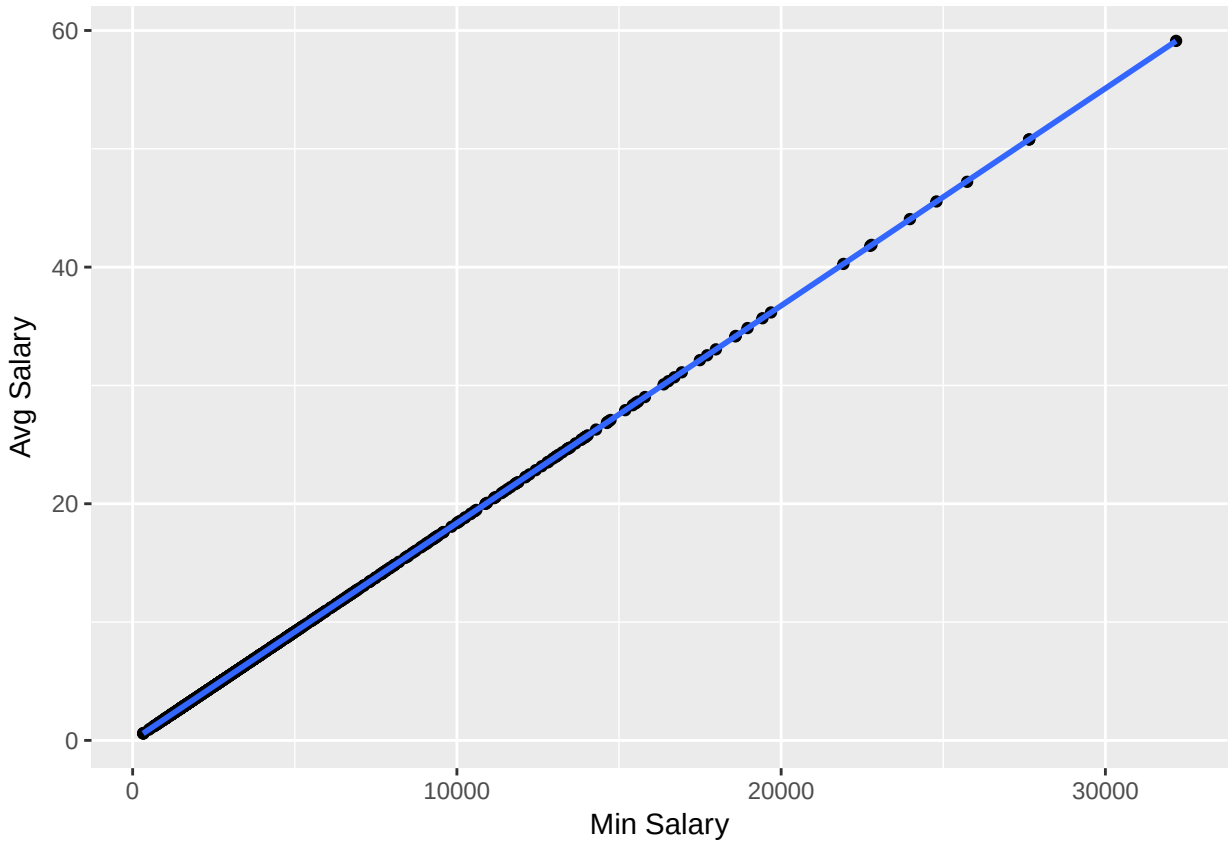
- **Min Salary:**

This variable 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 ilustration 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 possibles 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
```

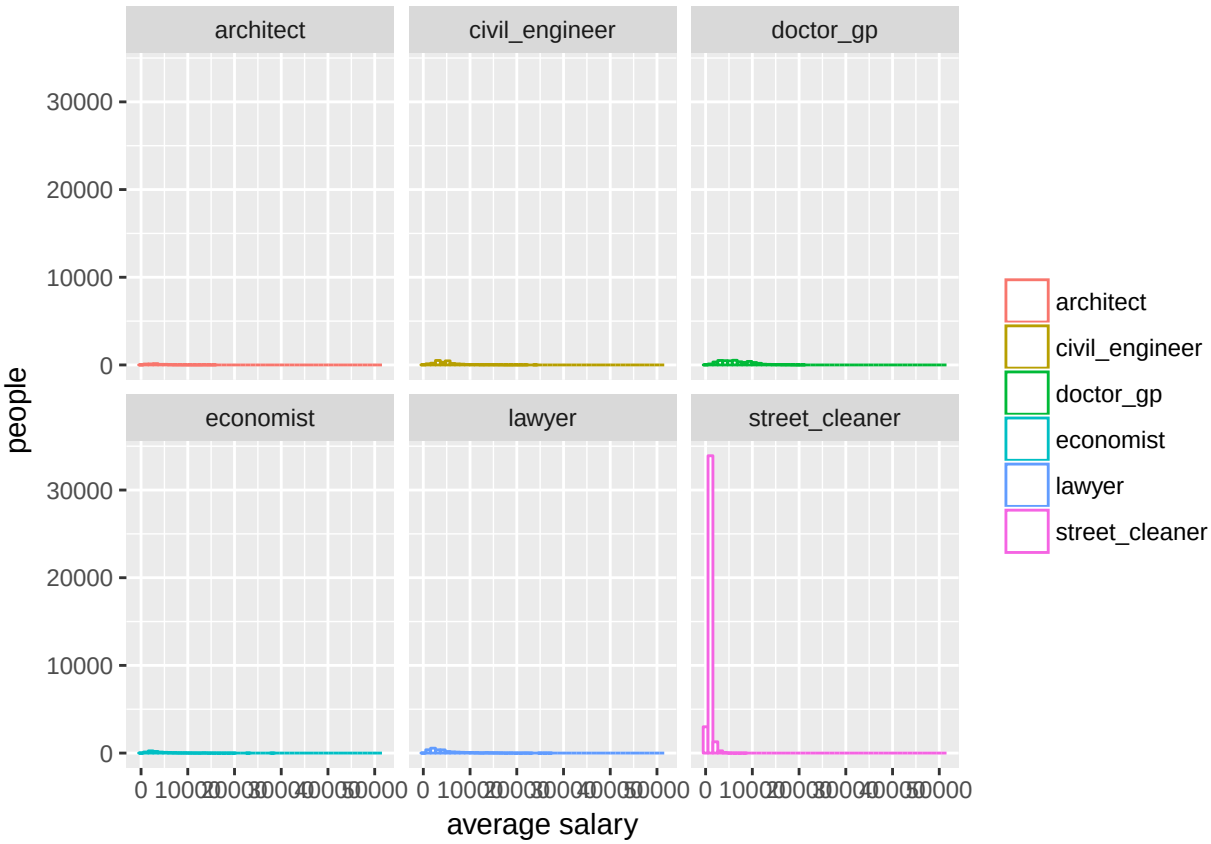
```
## `geom_smooth()` using method = 'loess'
```



- **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$CBO2002)

plot = ggplot(data = dataset, aes(dataset$avg_salary, color = dataset$CBO2002)) +
  geom_histogram(fill = "white", bins = 50) + labs(x= "average salary", y = "people") +
  facet_wrap(~ dataset$CBO2002) + 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 charging the visualization. Therefore, in the next section I start discussing how to scales 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 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 theses people are working more than the others! Therefore, from now on I'm going to use the average salary divised 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 natural logarithm on it, since it does not change the values order, just scales it.

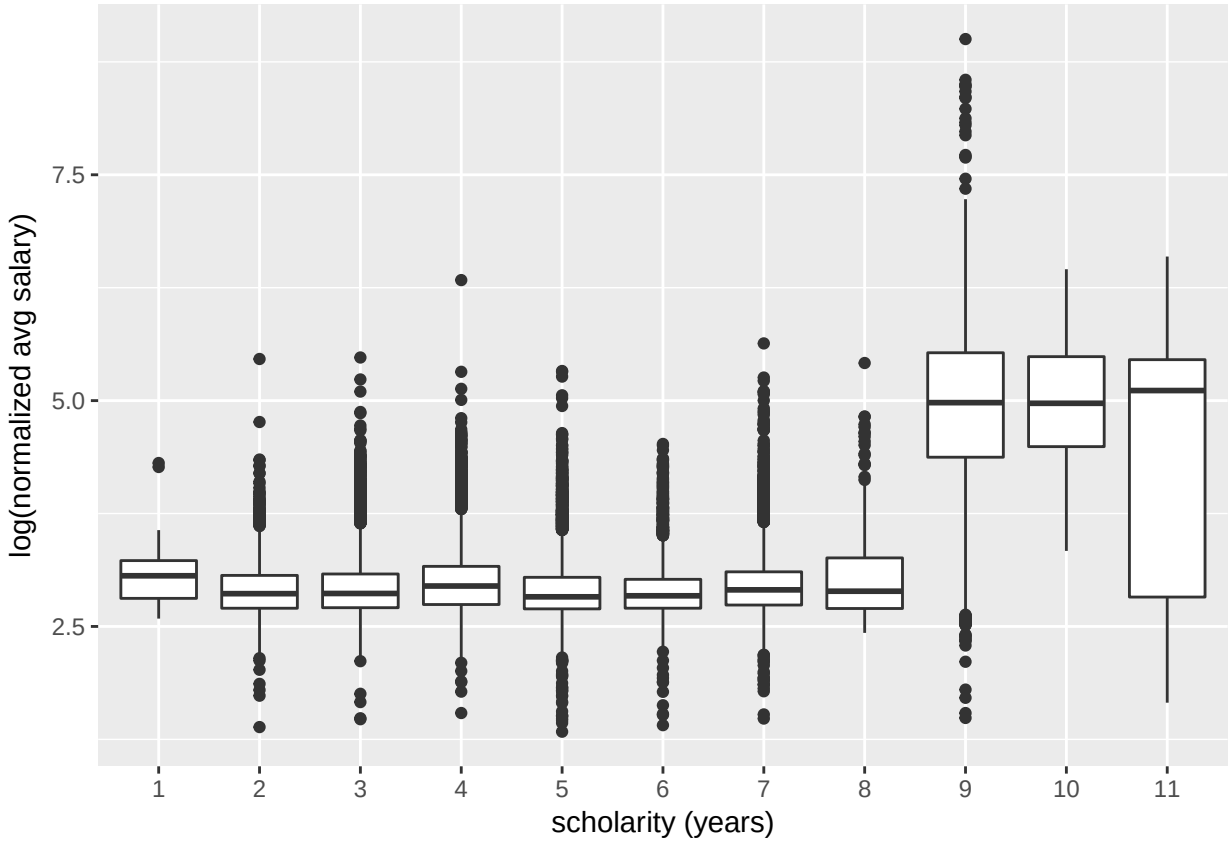
It is important to consider that applying the log operation on the average salary will change the interpretation of the results for the final regression, and it will be interpreted as the percentage salary variance given the absolute variance for the other variables.

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 emsemble of professions based solely on their normalized average

salary and their scholarity.

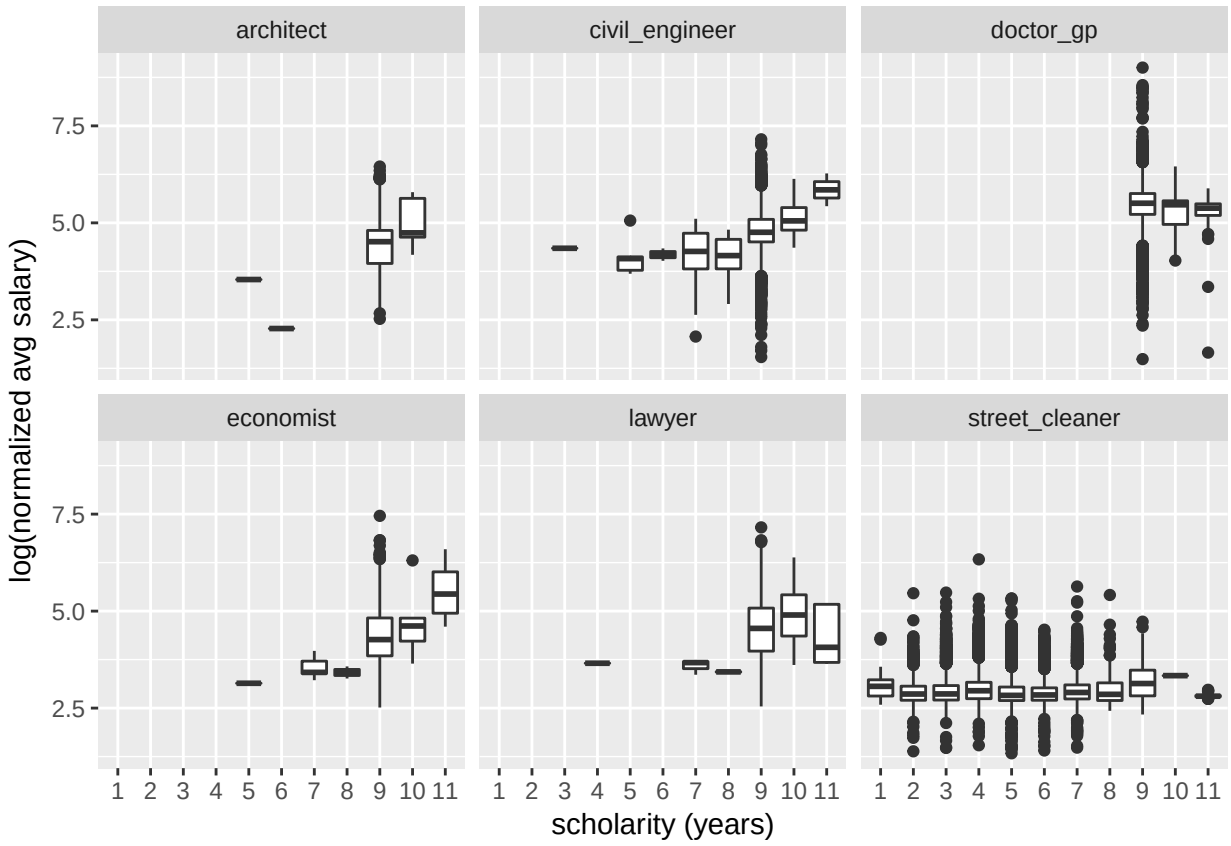
```
plot = ggplot(data= dataset, aes(factor(dataset$Scholarity), log(dataset$norm_avg_salary))) +
  geom_boxplot() + labs(x = "scholarity (years)", y = "log(normalized avg salary)")
plot
```



This visualization help us to identify a possible relationship between the salary and the scholarity. Although between the first seven years of scholarity we do not seem to have a considerable difference, from the eighth to the nineth 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 scholarity to eight, and a different set of professions that require at least nine years of scholarity. However, fortunately, after plotting scholarity 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.

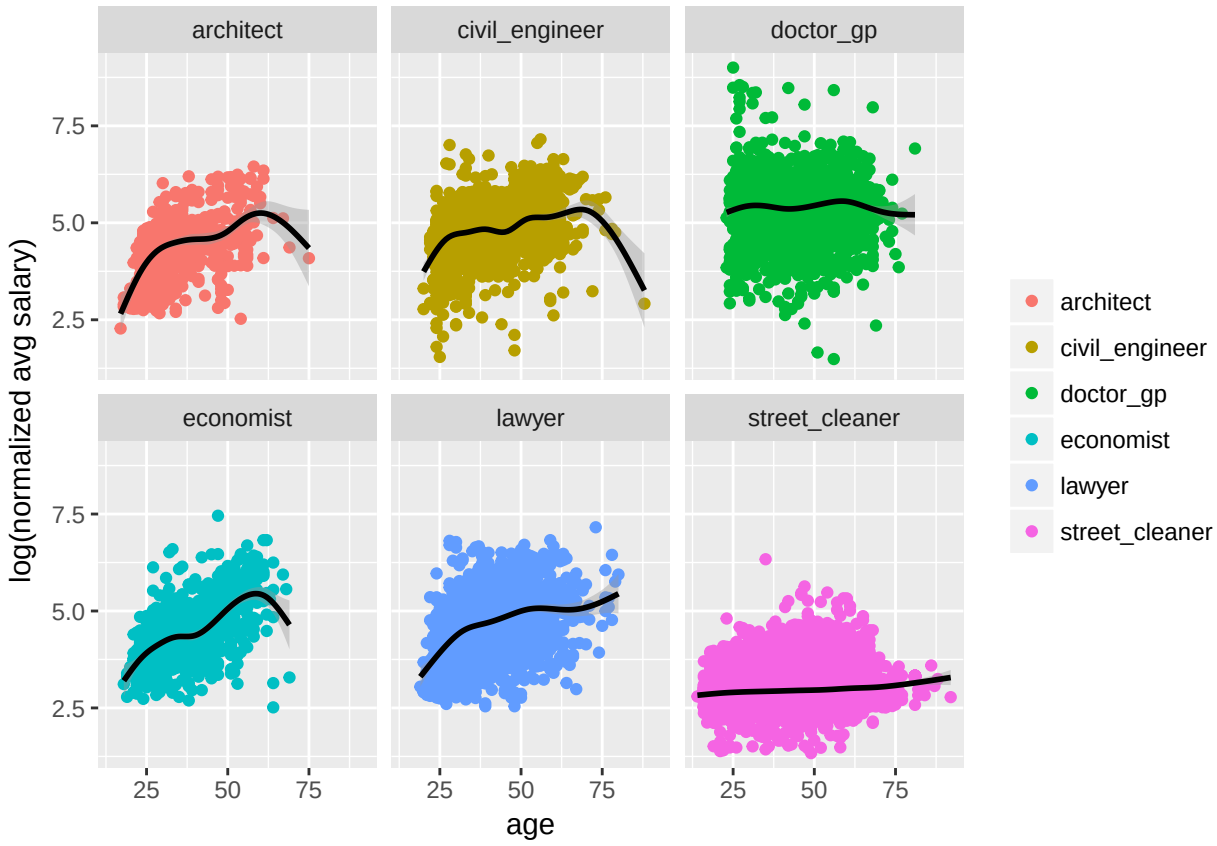
```
plot = ggplot(data= dataset, aes(factor(dataset$Scholarity), log(dataset$norm_avg_salary))) +
  geom_boxplot() + labs(x = "scholarity (years)", y = "log(normalized avg salary)") +
  facet_wrap(~ dataset$CBO2002)
plot
```



Age

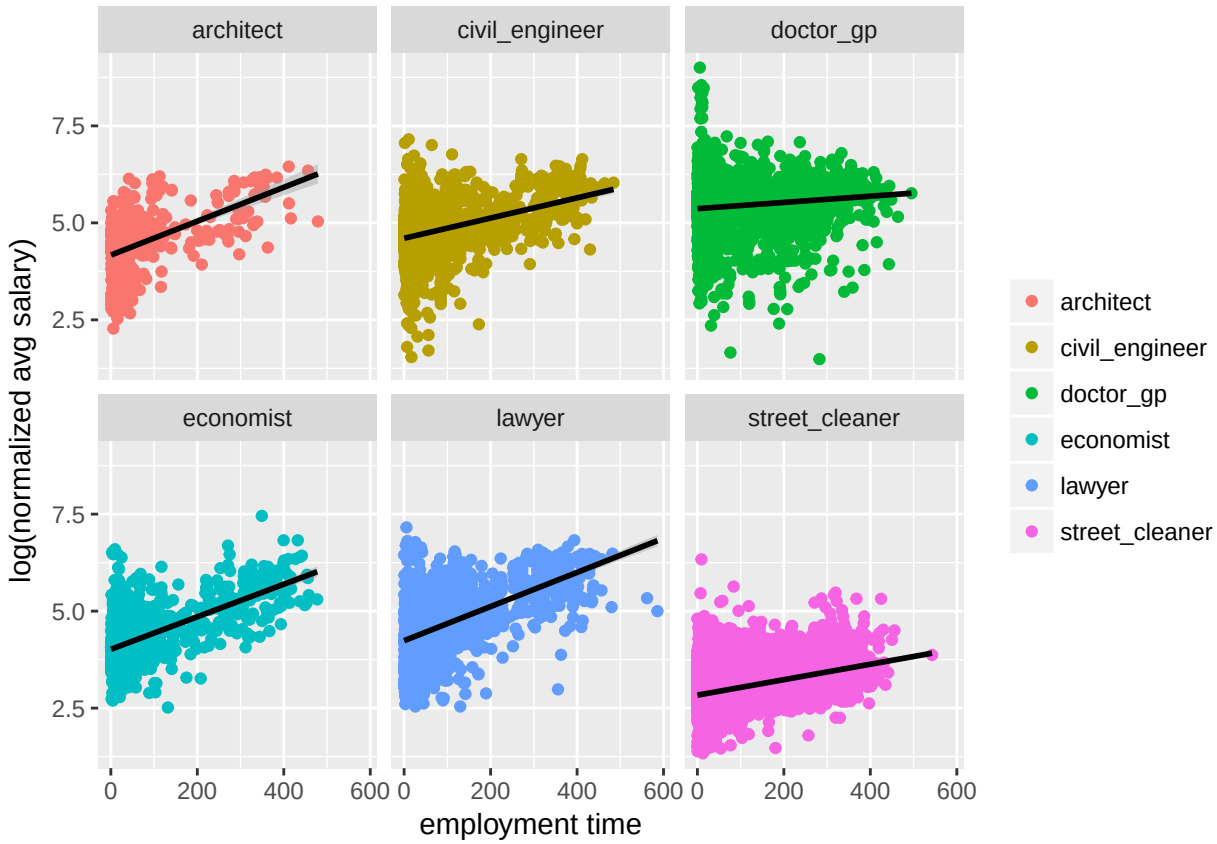
```
plot = ggplot(data = dataset, aes(dataset$age, log(dataset$norm_avg_salary),
                                  color = dataset$CBO2002)) + geom_point() +
  geom_smooth(color = "black") + labs(x = "age", y = "log(normalized avg salary)") +
  facet_wrap(~ dataset$CBO2002) + theme(legend.title=element_blank())
plot
```

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



Employment Time

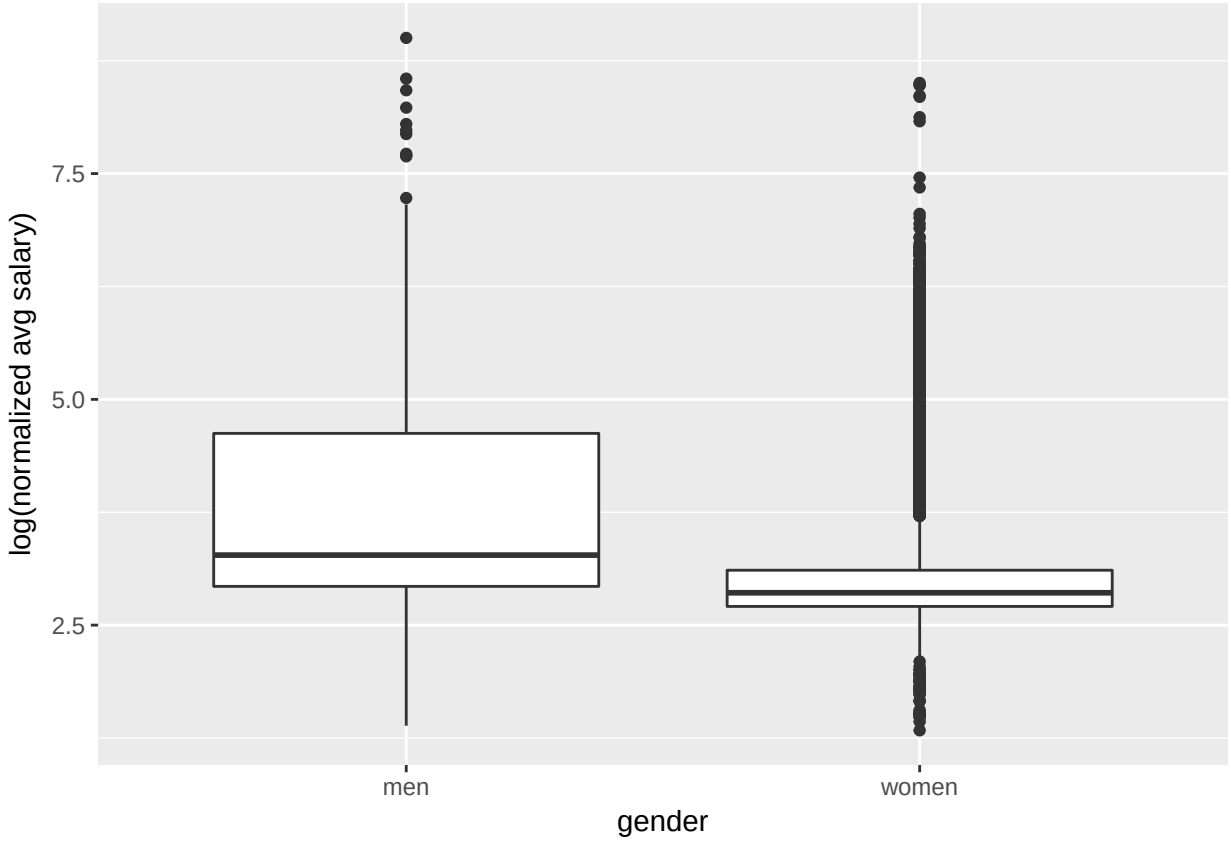
```
plot = ggplot(data = dataset, aes(dataset$employment_time, log(dataset$norm_avg_salary),
                                  color = dataset$CBO2002)) + geom_point() +
  geom_smooth(color = "black", method = lm) + labs(x = "employment time",
                                    y = "log(normalized avg salary)") +
  facet_wrap(~ dataset$CBO2002) + theme(legend.title=element_blank())
plot
```



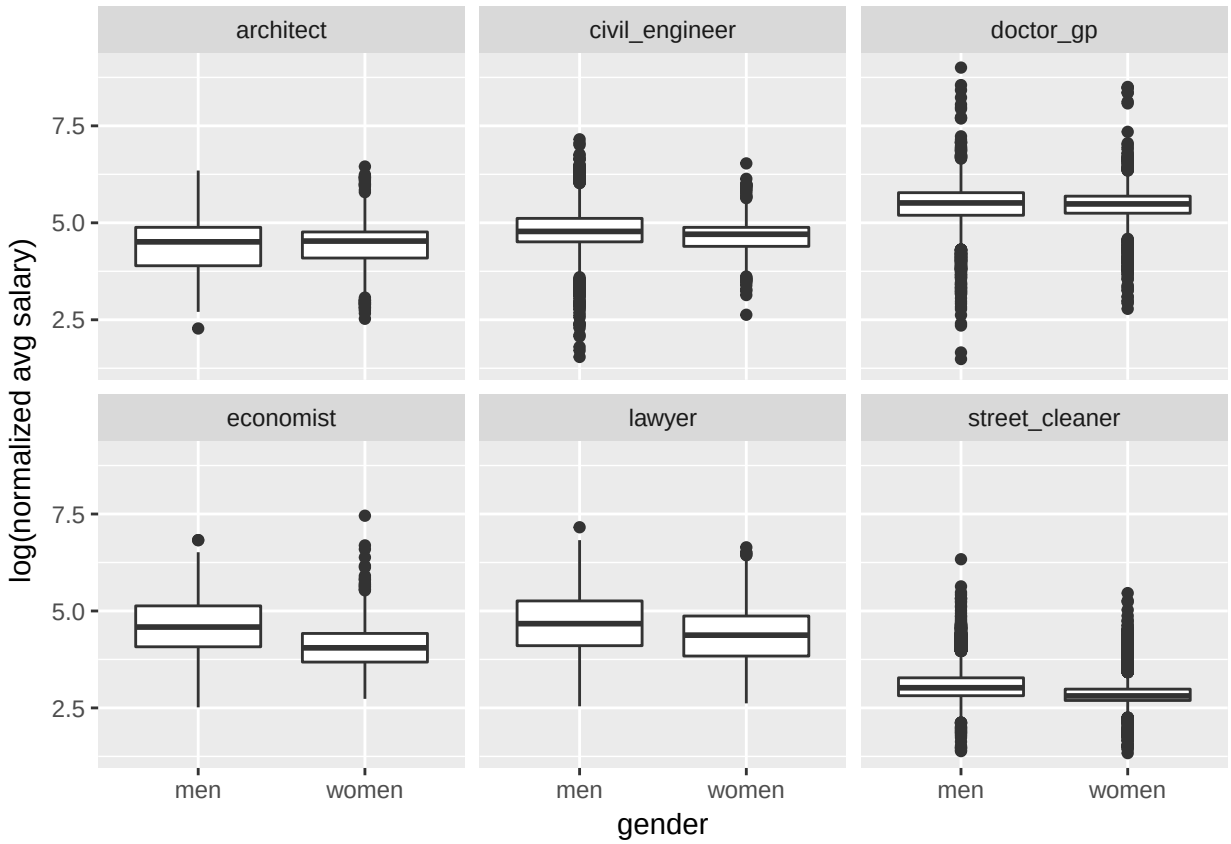
Similar as when comparing age with employment time, the observations seem to have a huge variability.

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$CBO2002)
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 norm_avg_salary = B0 + B1 scholarity + B2 age + B3 age^2 + + B4 gender + B5 employment_time + error

Where B0 is the salary of an individual with no education and no experience, and B1, B2, B3, B4 and B5 are the adjust parameters.

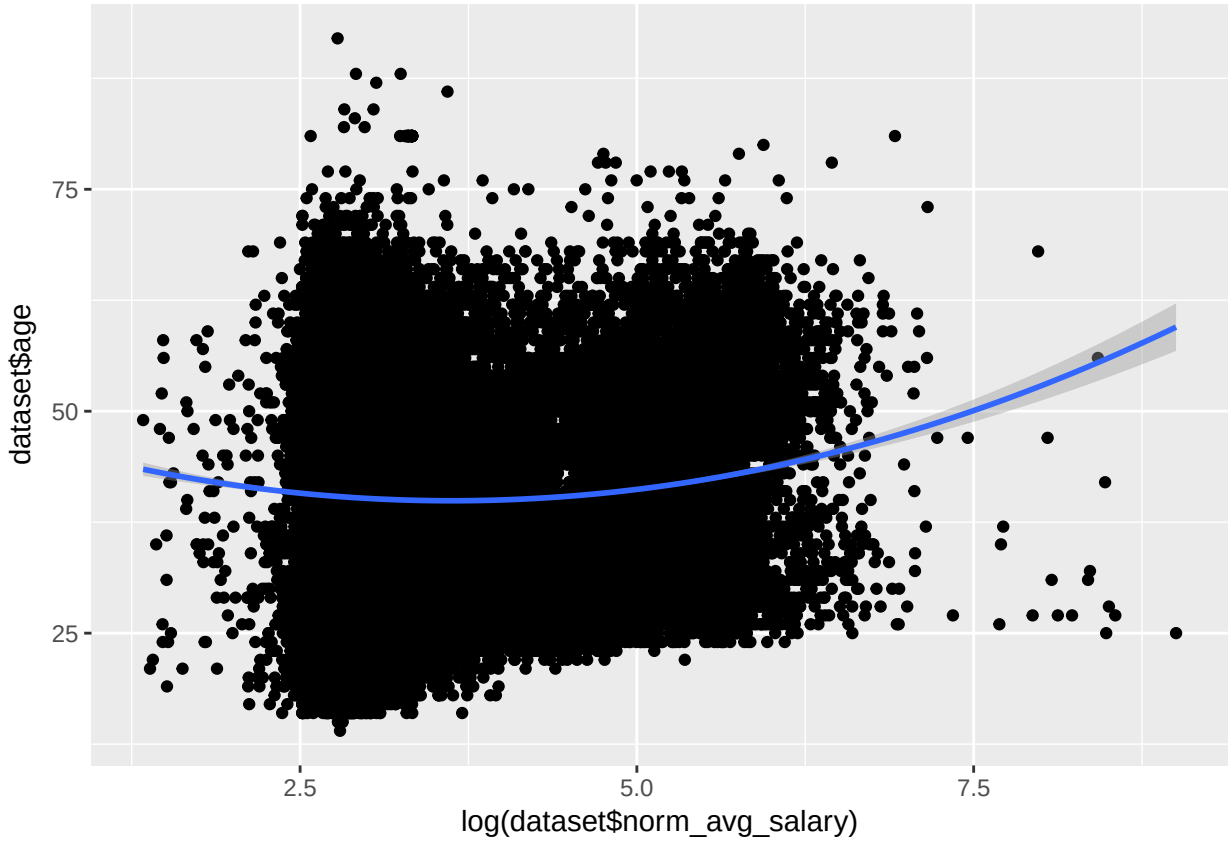
```
reg = lm(data = dataset,
         formula = log(dataset$norm_avg_salary) ~ dataset$Scholarity +
           dataset$age + dataset$gender + dataset$employment_time  +
           I(dataset$age^2))

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, discarting 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, which was illustrated in the last section and 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 good R-squared value!

Also, a pitfall that I fell when 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 (B0 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).

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").

References

- [1] Sites aboutt Mincer Earnings Function:

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

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

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

- [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