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).

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

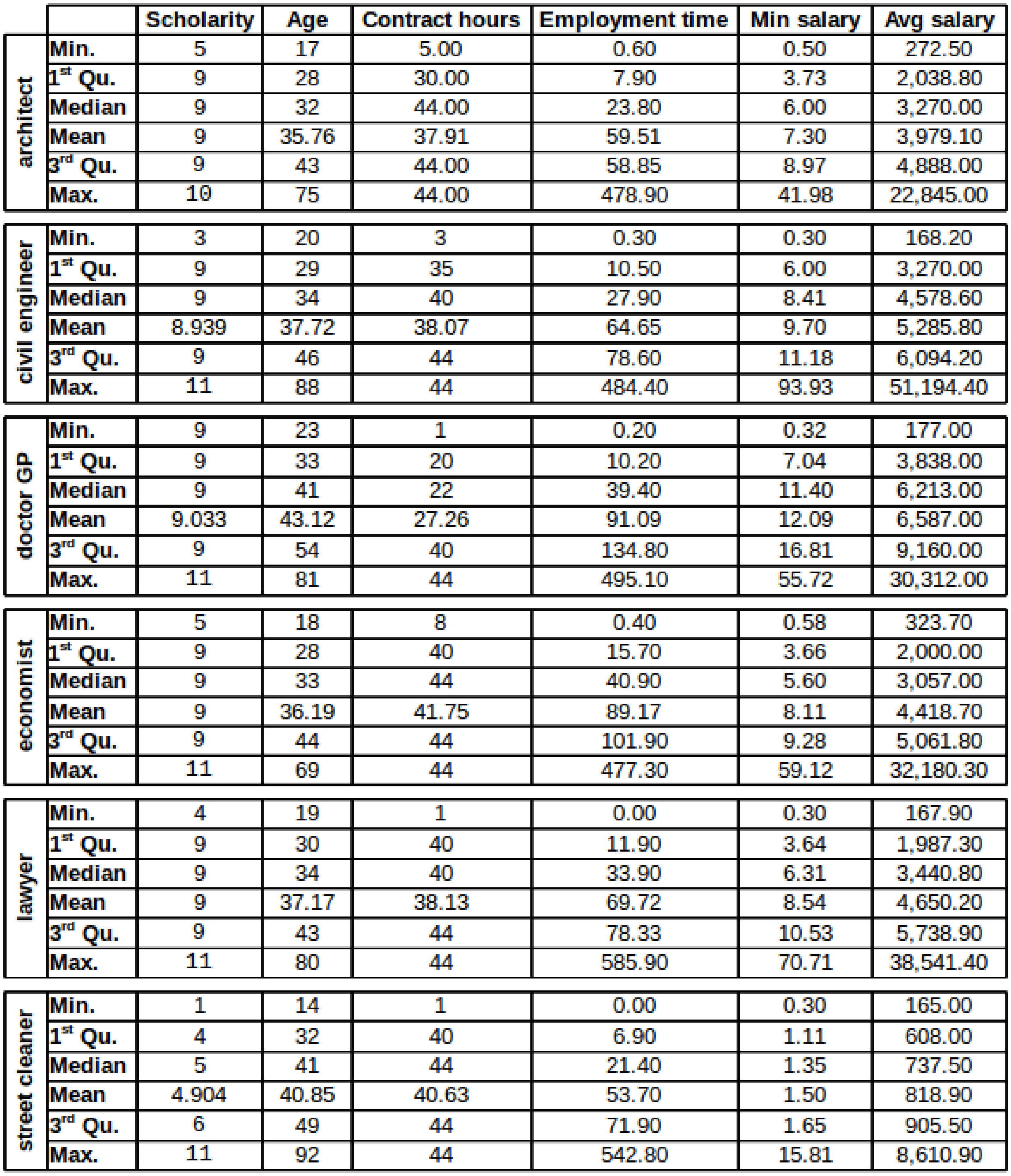
Economist: 961

Lawyer: 2.476

Street Cleaner: 49.001

General Analysis for each variable:

- **Scholarity:**



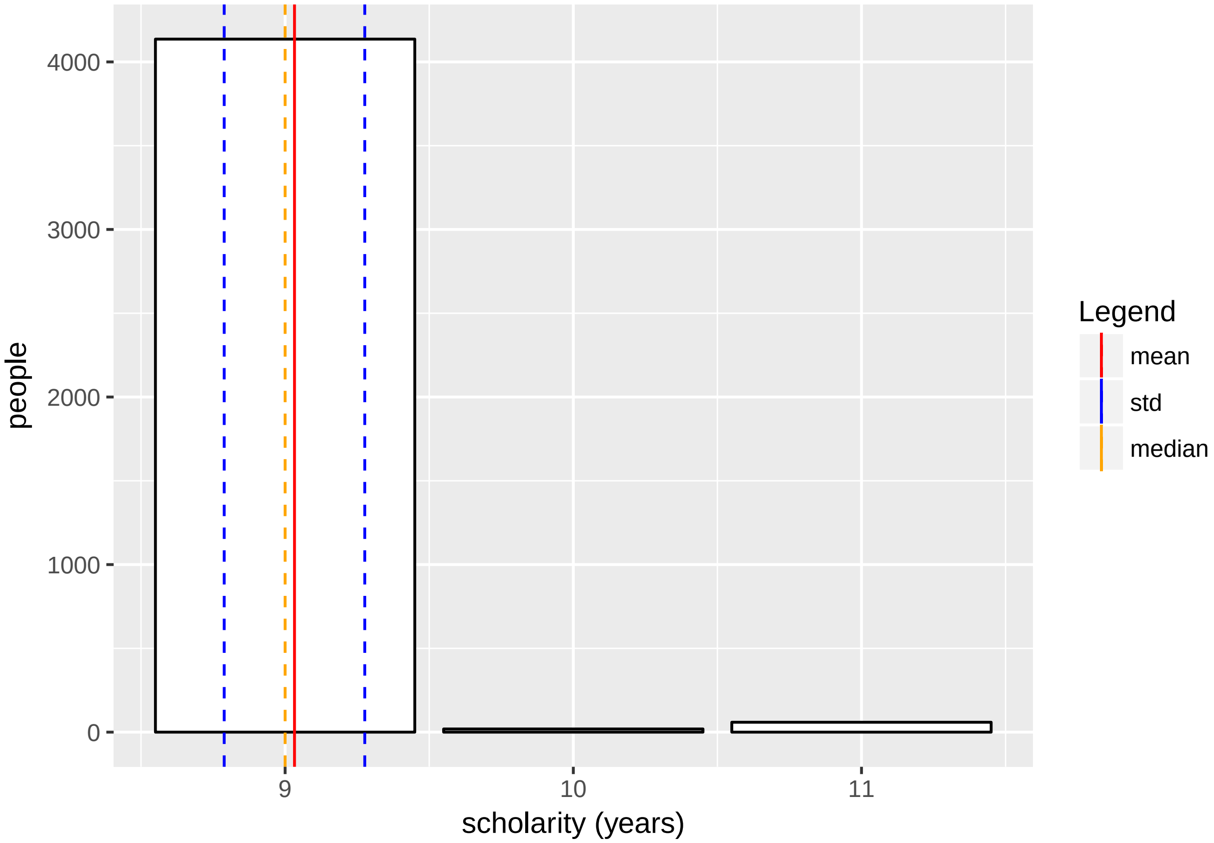
		Scholarity	Age	Contract hours	Employment time	Min salary	Avg salary
architect	Min.	5	17	5.00	0.60	0.50	272.50
	1st Qu.	9	28	30.00	7.90	3.73	2,038.80
	Median	9	32	44.00	23.80	6.00	3,270.00
	Mean	9	35.76	37.91	59.51	7.30	3,979.10
	3rd Qu.	9	43	44.00	58.85	8.97	4,888.00
	Max.	10	75	44.00	478.90	41.98	22,845.00
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:

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".

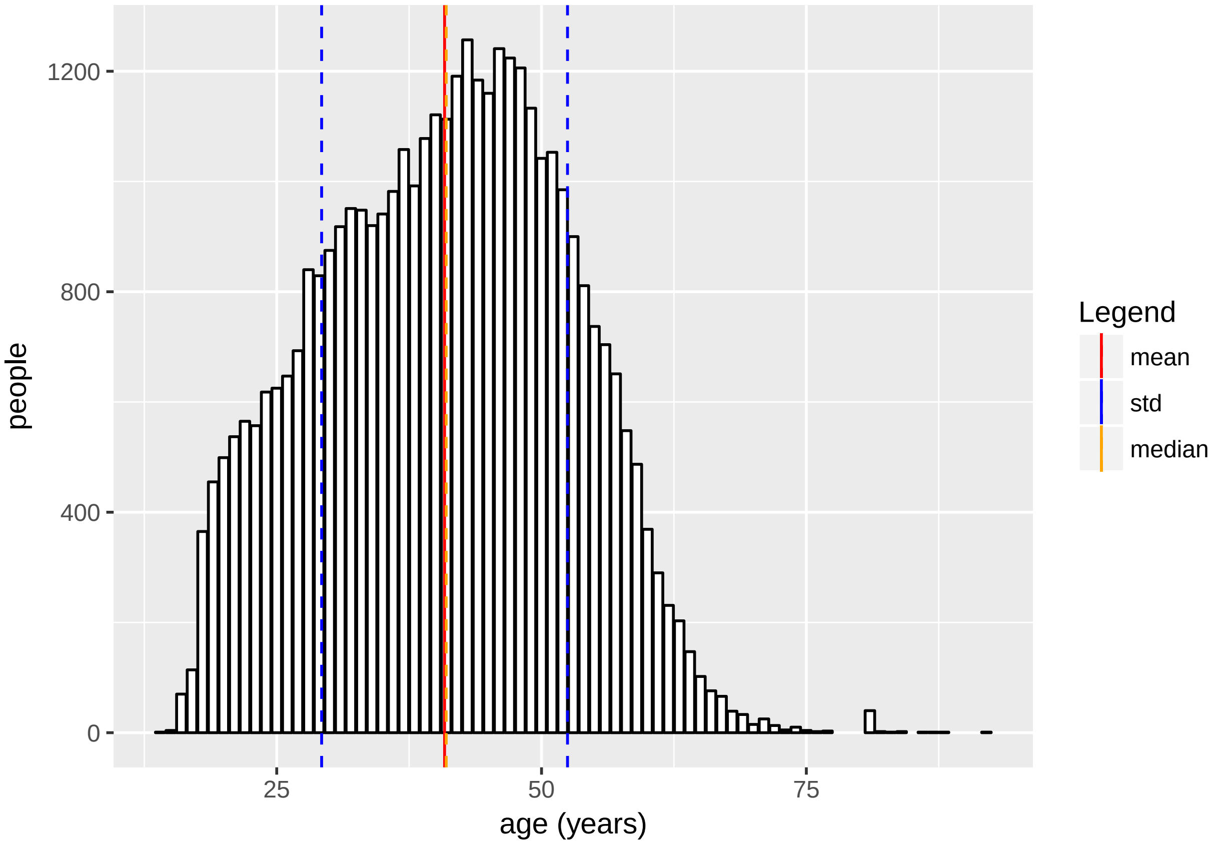
```
library(ggplot2)
load(file="data/economist.Rdata")
load(file="data/street_cleaner.Rdata")
load(file="data/doctor_general_practice.Rdata")
meanE <- mean(doctor_general_practice$Scholarity)
std <- sd(doctor_general_practice$Scholarity)
plot = ggplot(data = doctor_general_practice, aes(doctor_general_practice$Scholarity))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



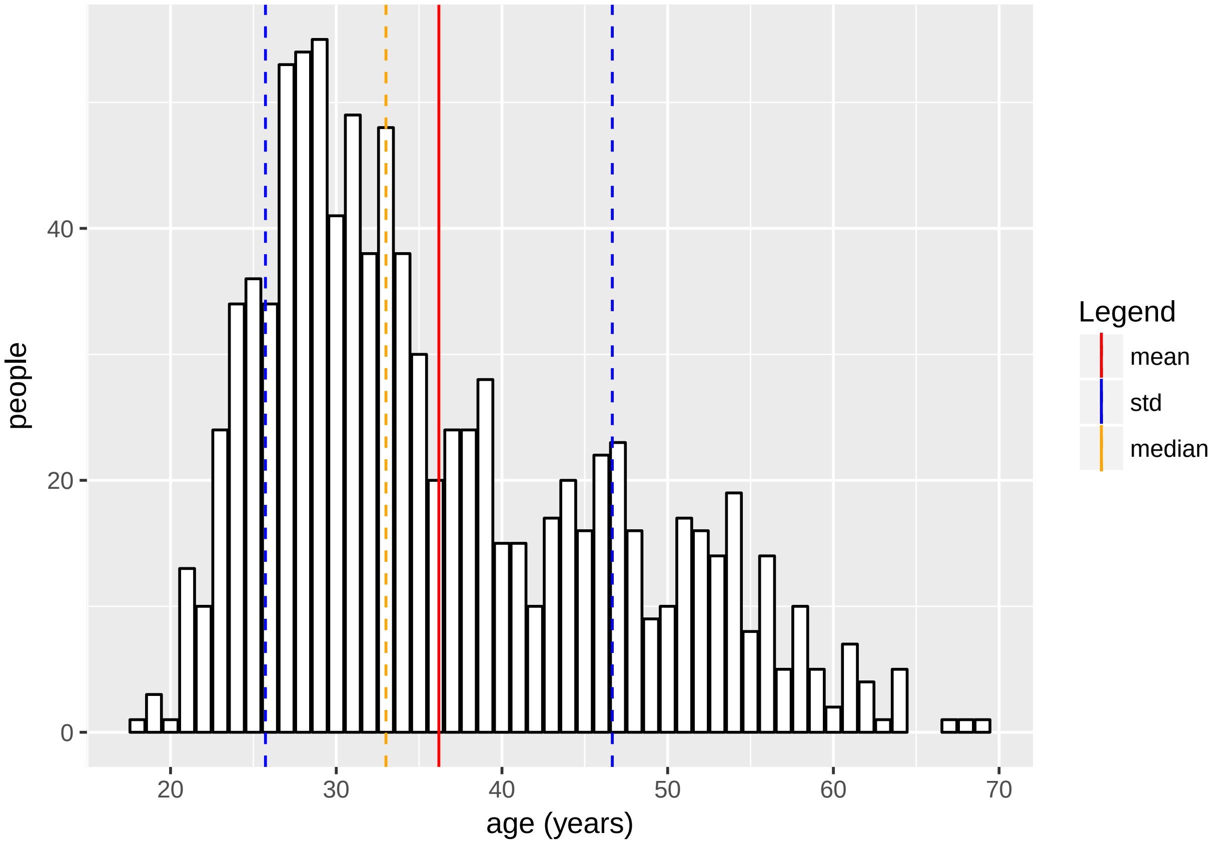
- **Age:**

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 observation on this age for the street cleaning dataset. However, for the other professions, it's expected from the employee to have more time of education in order to fulfill their tasks, and thus we have a higher minimum for the other 5 datasets, as well as a higher number of registerment around the 25 years, a common age to finish college studies.

```
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
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
economistPlot
```



- **Contract Hours**
- **Employment Time**
- **Min Salary**
- **Average Salary**

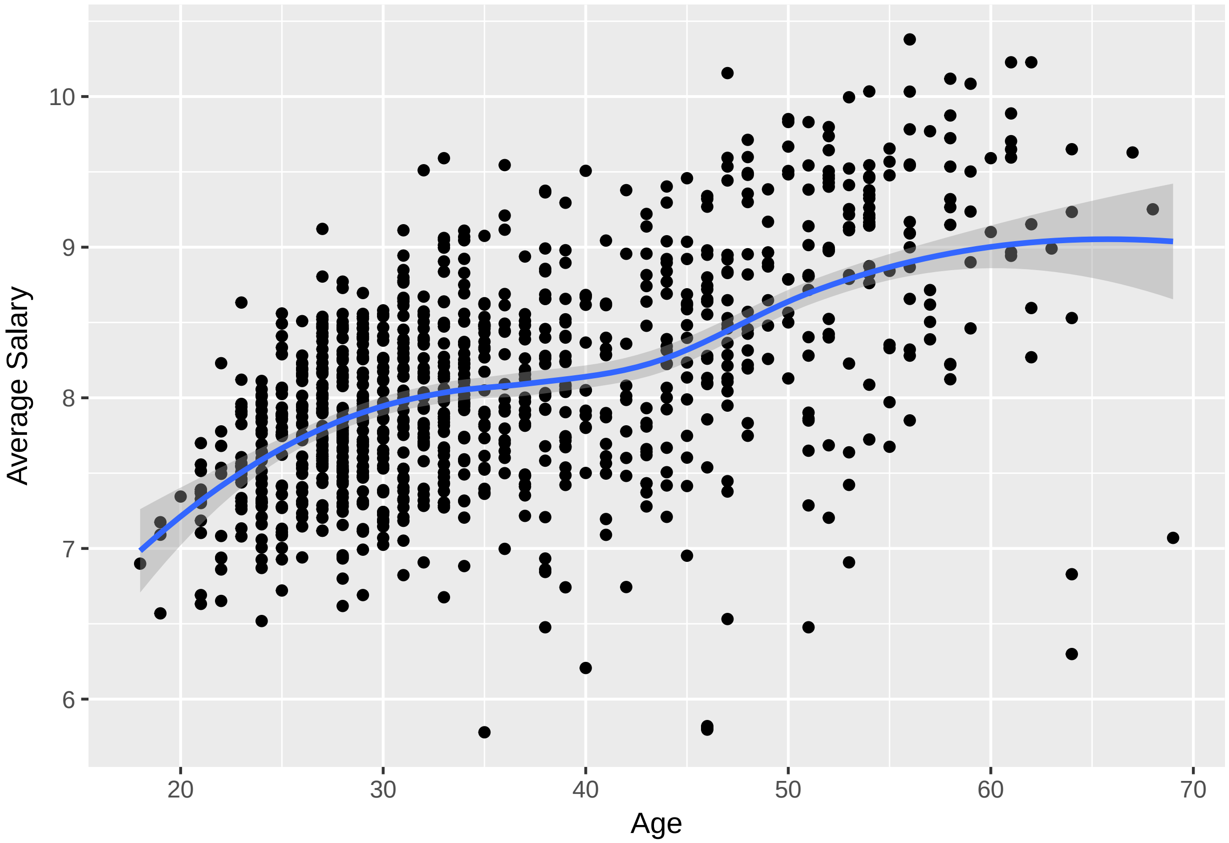
Variables of interest against Average Salary

For these plots, we are using the log of the average salary. This change does not affect the values order, it only escales the data, helping to visualize it. In this section we are focusing in age and gender.

Age x Salary

```
ggplot(data = economist, aes(economist$age, log(economist$avg_salary))) + geom_point() + labs(y = "Avera
```

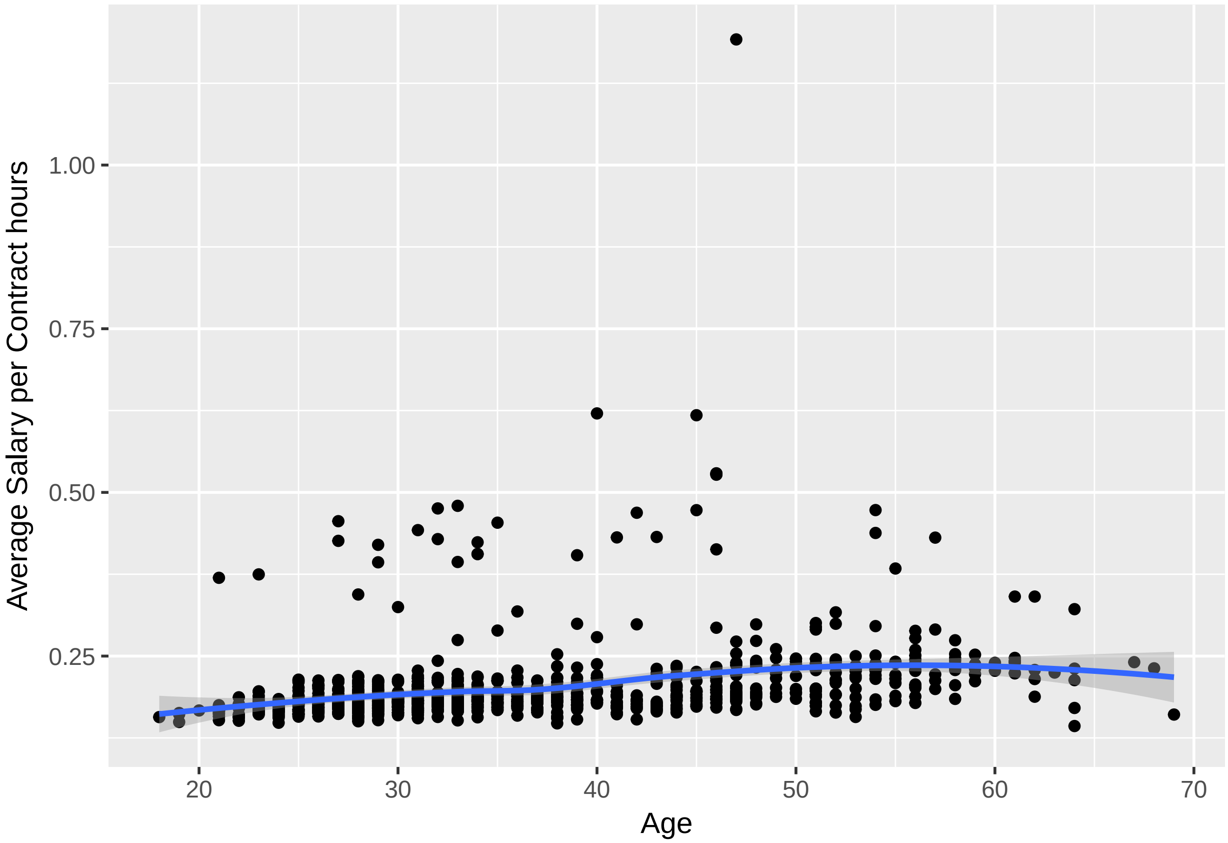
```
## `geom_smooth()` using method = 'loess'
```



Looking at these values together, it gives the impression that the average salary (despite some outliers) is increasing together with age. However, here we are not considering the amount of hours that each economist is working per month, and that could give us the false impression that someone old is receiving a lot more money (or vice-versa), where in reality they could be just working more. To fix that, I tried to plot the data using the variable contract hours to divide the average salary value, resulting in the plot below.

```r
qplot(economist$age, (log(economist$avg_salary) / economist$contract_hours), ylab = "Average Salary per
```

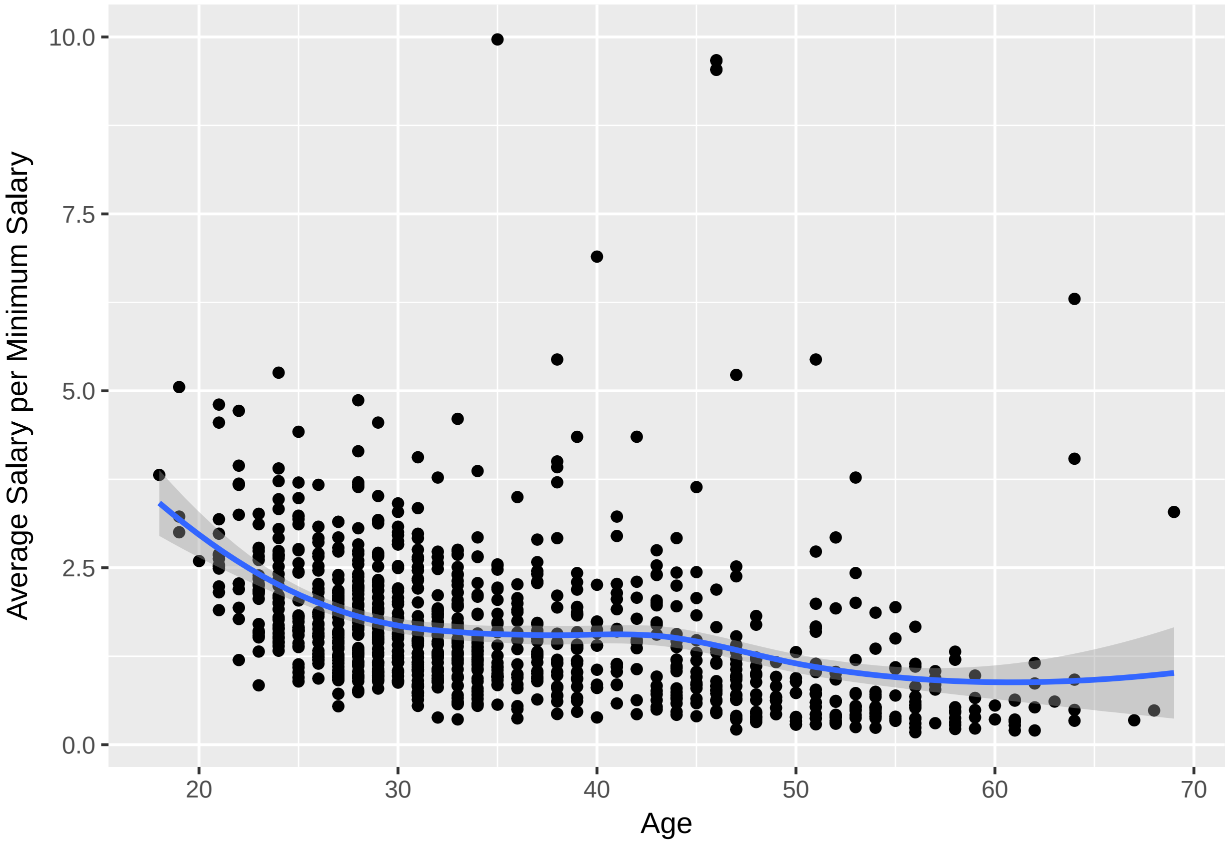
```
## `geom_smooth()` using method = 'loess'
```



After dividing the average salary per number of hours, the values are way more uniform.

```
qplot(economist$age, (log(economist$avg_salary) / economist$min_salary), ylab = "Average Salary per Min
```

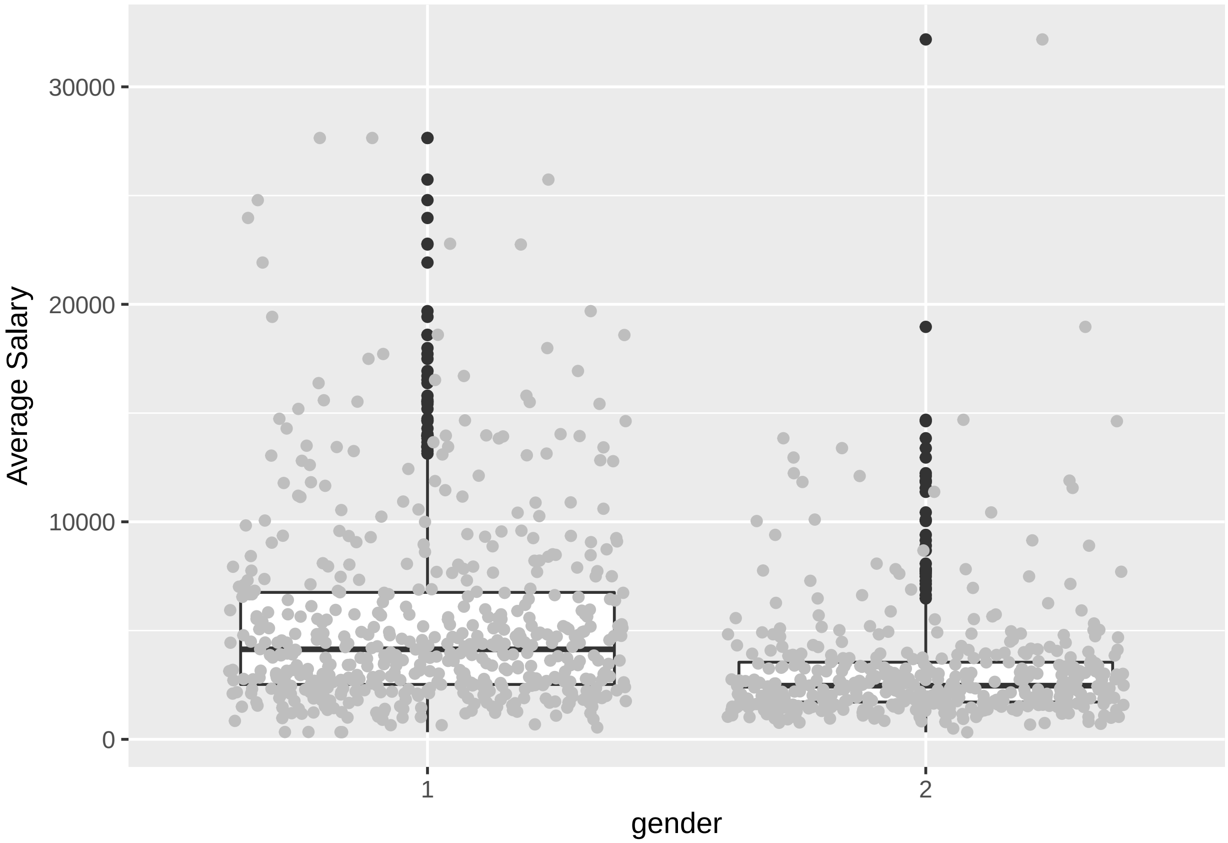
```
## `geom_smooth()` using method = 'loess'
```



????????????????

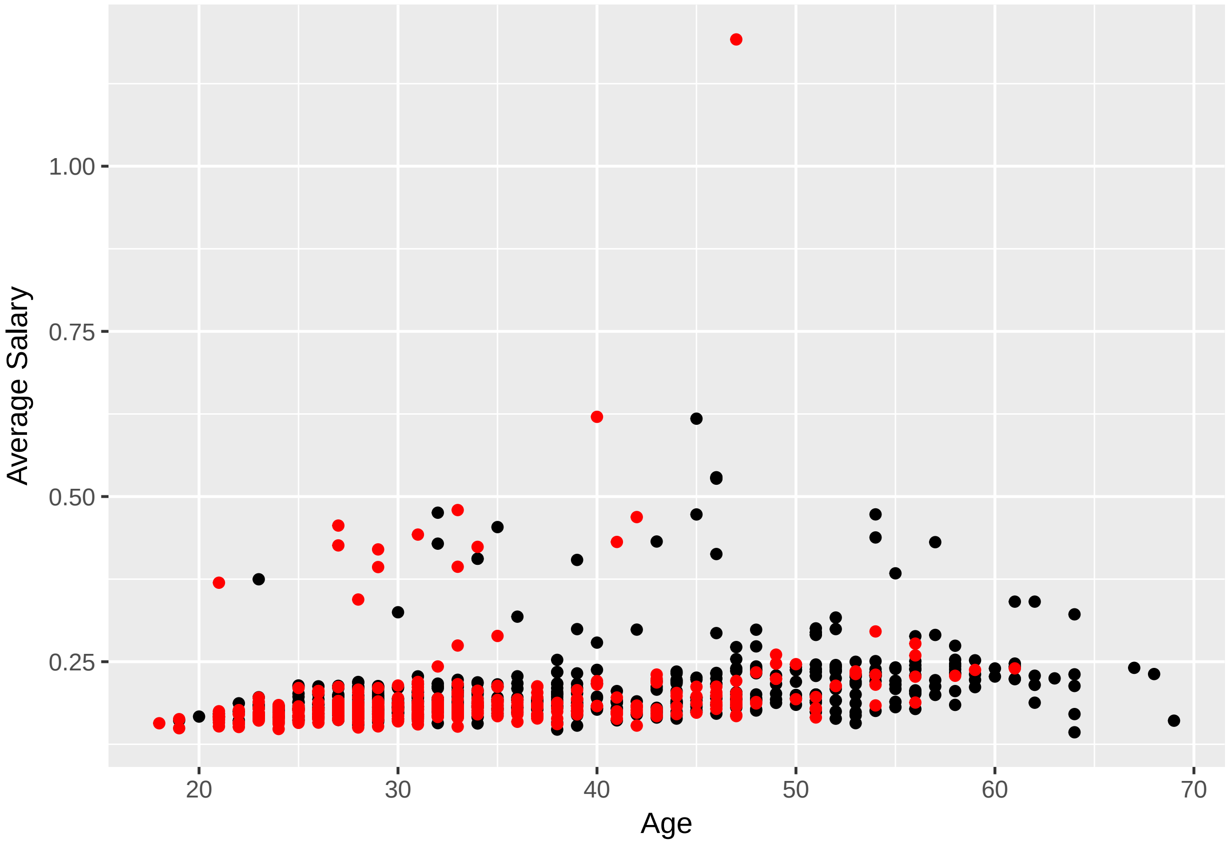
gender x salary

```
ggplot(data=economist, aes(x= factor(economist$gender), y=economist$avg_salary)) + geom_boxplot() + labs
```



Gender and Age against Average Salary

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



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").