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. These datasets contain information about people registered as regular workers, following the "CBO" (brazilian official classification of professions). 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").

We have six datasets, each one representing a different profession: architecture, medicine, engineering, economy, law and street cleaning, and they present information about number of hours, salary average, the salary minimum per hour, the gender, the age and employment time.

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 gender discrimination? 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?

First Dataset: Economist set

We have seven variables in this dataset: contract hours, age, gender, scholarity, average salary, minimum per hour, and employment time. The table bellow was generated using R "summary" command and sit shows the general

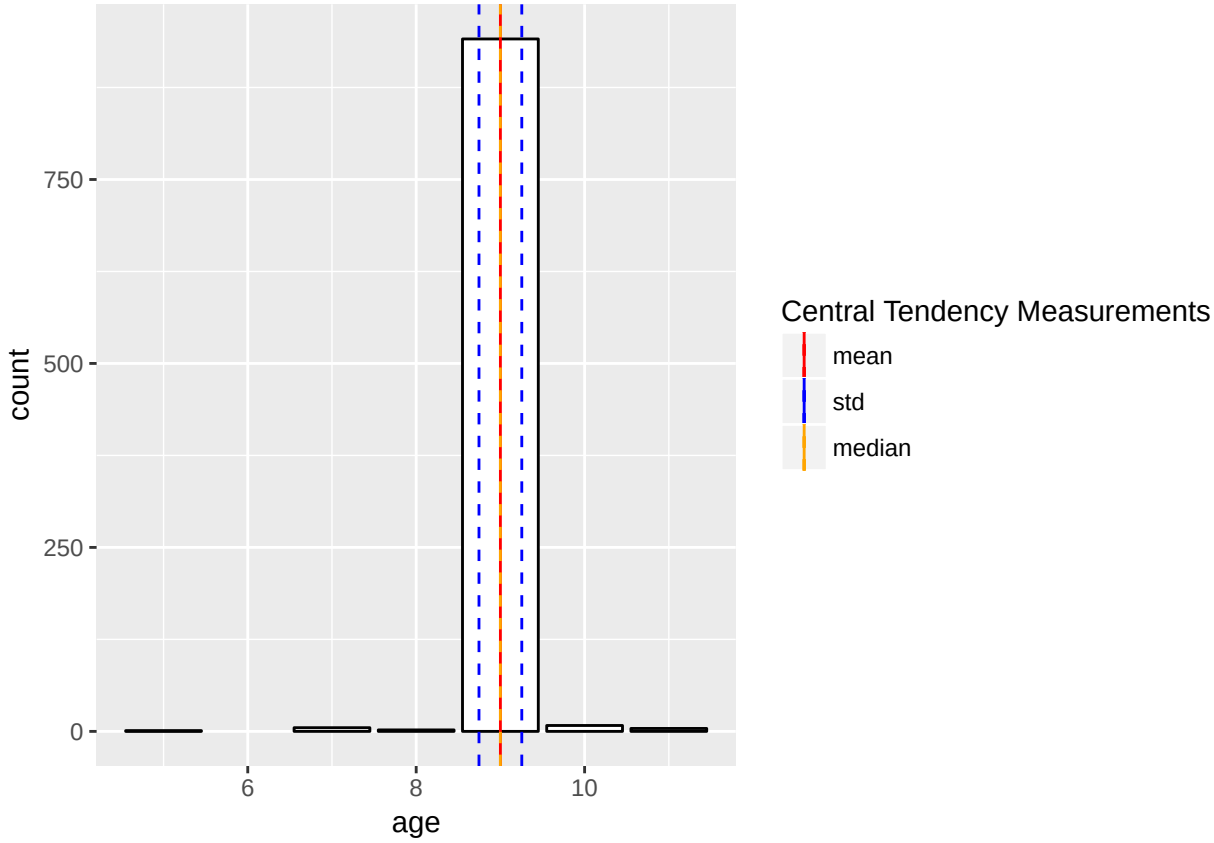
	scholarity	age	gender	contract_hours	employment_time	min_salary	avg_salary
Min.	5	18	1	8	0.4	0.58	323.7
1st Qu.	9	28	1	40	15.7	3.66	2000
Median	9	33	1	44	40.9	5.6	3057
Mean	9	36.19	1.461	41.75	89.17	8.109	4418.7
3rd Qu.	9	44	2	44	101.9	9.28	5061.8
Max.	11	69	2	44	477.3	59.12	32180.3

Figure 1: General metrics for the variables in the dataset

- Scholarity:

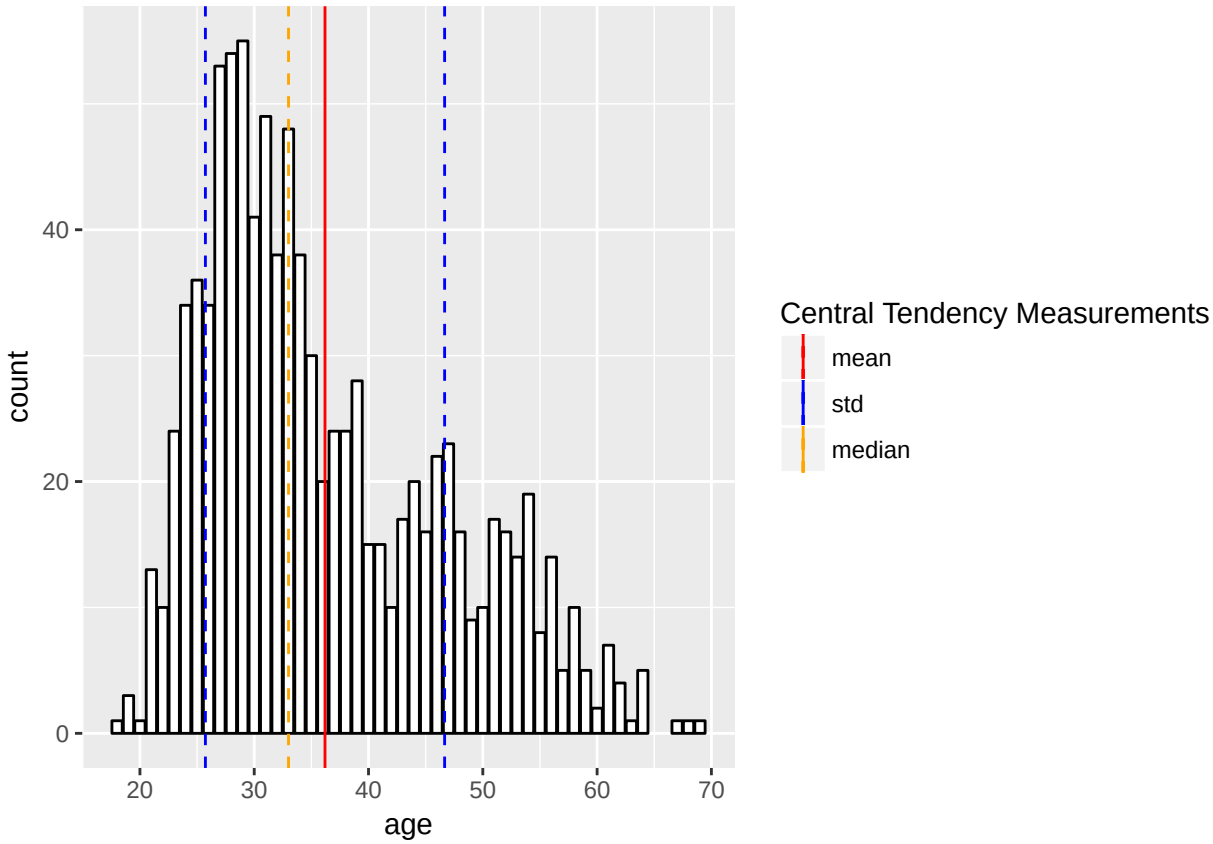
Looking in how concentrate are the values for scholarity, I concluded that it would be difficult to infer information considering different scholarities inside a profession. However, probably it would be interesting to see how it behaves comparing different professions.

```
library(ggplot2)
load(file="data/economist.Rdata")
meanE <- mean(economist$Scholarity)
std <- sd(economist$Scholarity)
ggplot(data = economist, aes(economist$Scholarity)) + geom_bar(fill="white", colour = "black") + labs(x=
```



- Age: For age, we have an interesting number of di

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



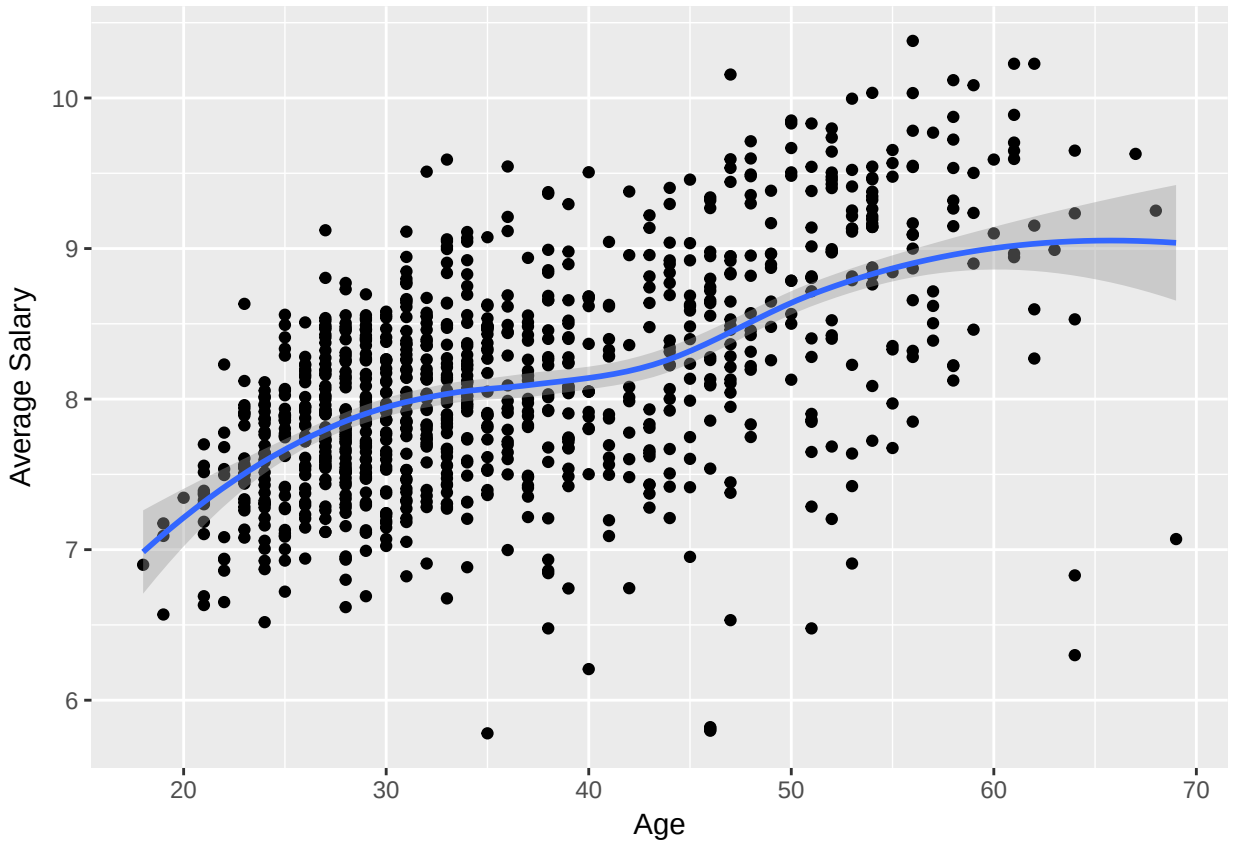
Variables of interest versus Average Salary

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

```
qplot(economist$age, log(economist$avg_salary), ylab = "Average Salary", xlab = "Age") + geom_smooth()
```

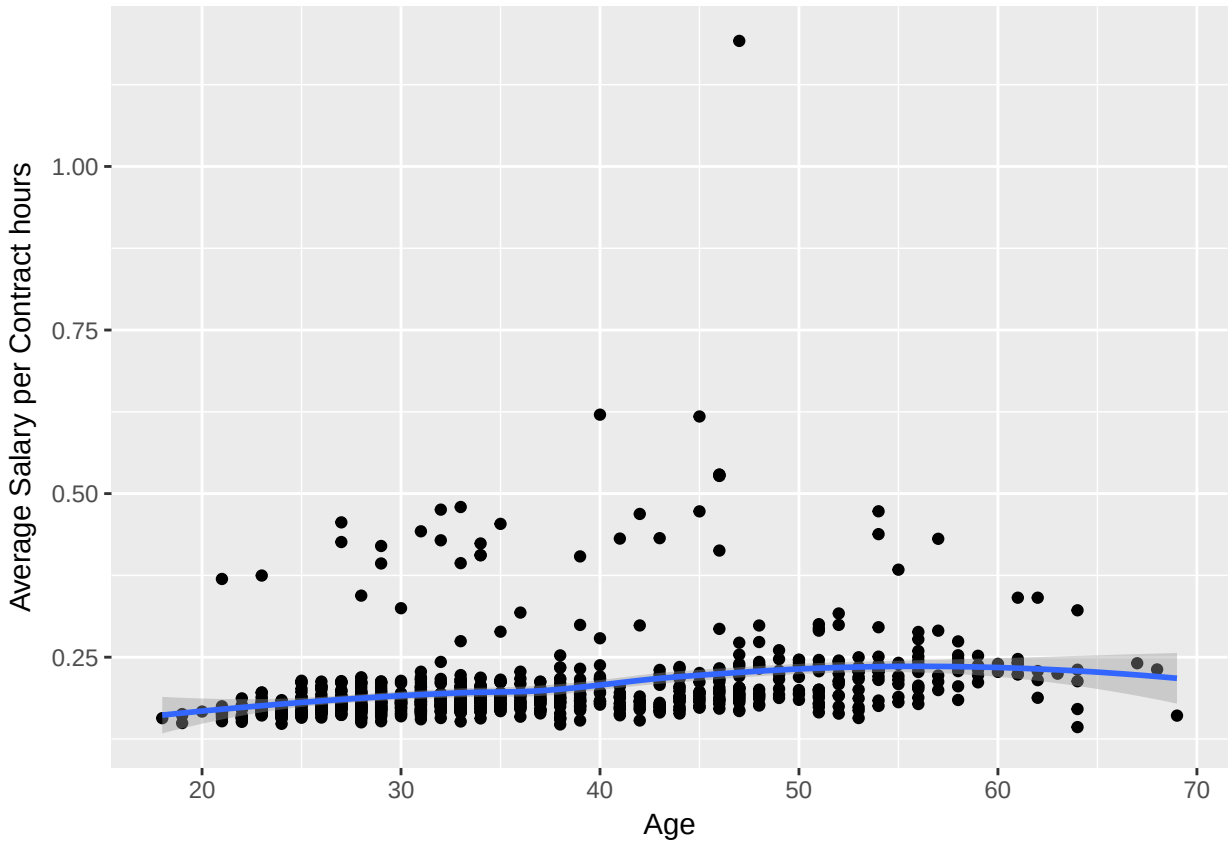
```
## `geom_smooth()` using method = 'loess'
```



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.

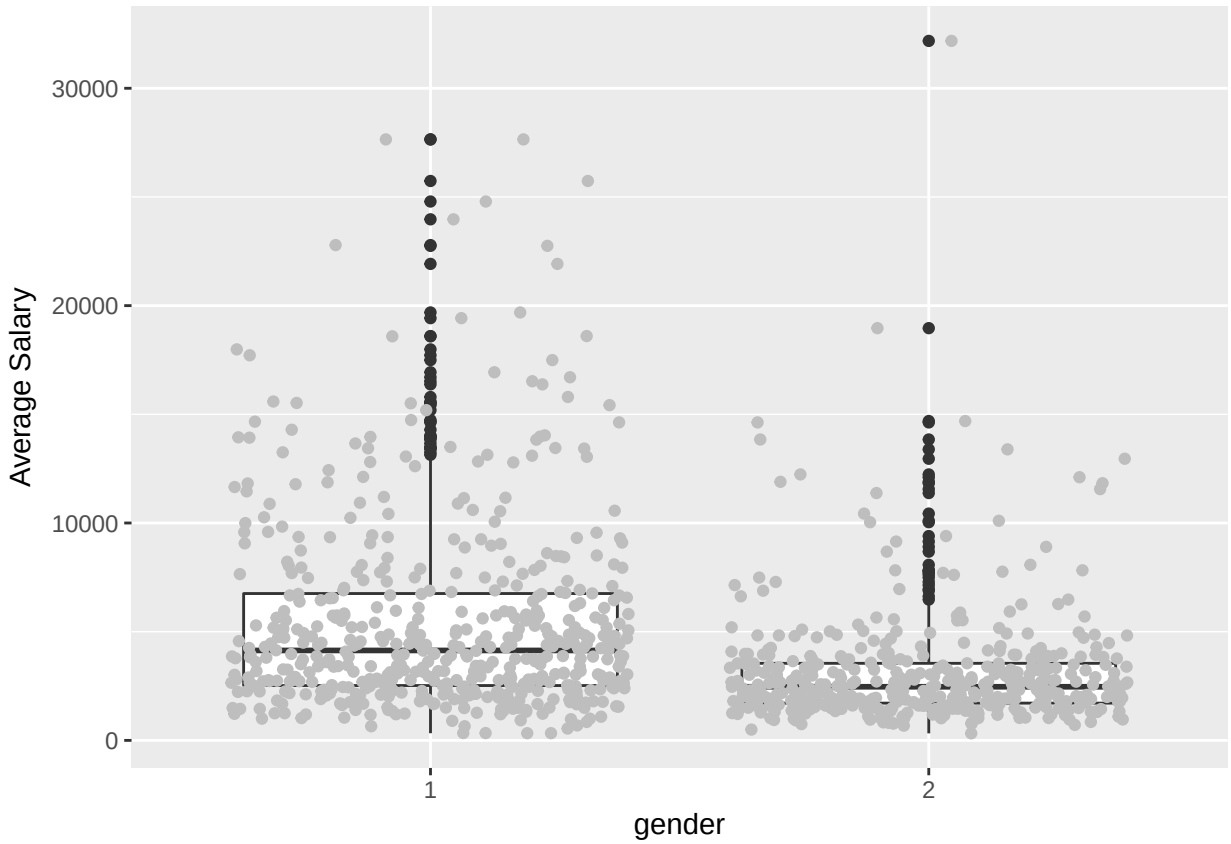
```
qplot(economist$age, (log(economist$avg_salary) / economist$contract_hours), ylab = "Average Salary per
```

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



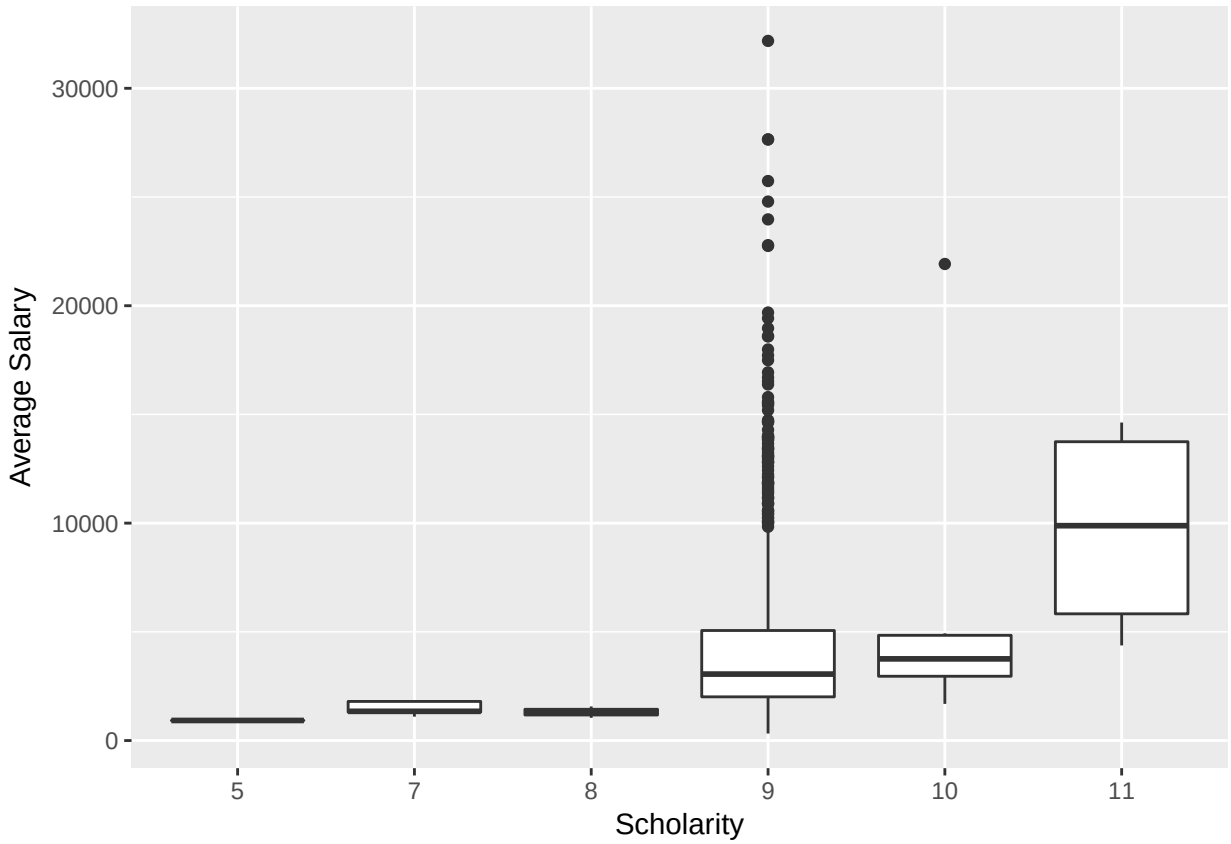
gender x salary

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



scholarity x salary

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


30000
20000
10000
0
Average Salary
5
7
8
9
10
11
Scholarity