A study about salary difference in Brazil

Marcely Zanon Boito

December 21, 2016

The Data sets

We have six data sets, each one representing a different working field: architecture, medicine, engineering, economy, law and street cleaning. All these datasets have information about people registred as payed professionals for that area, and they have information about number of hours, salary average, the salary minimum per hour, the gender, the age and employment time.

The Problem and the Hypothesis:

Having information about different six different working fields at Brazil, we want to identify:

1. Where we have the bigger salary gaps per gender;
2. How time of scolarity affects:

- 2.1. The salary;
- 2.2. The employment time;

3. The bigger estability (measured as average employment time);
4. How age affect the salary:

- 4.1. When a person gets older in Brazil, do they gain more or less money?

First Dataset: Economist set

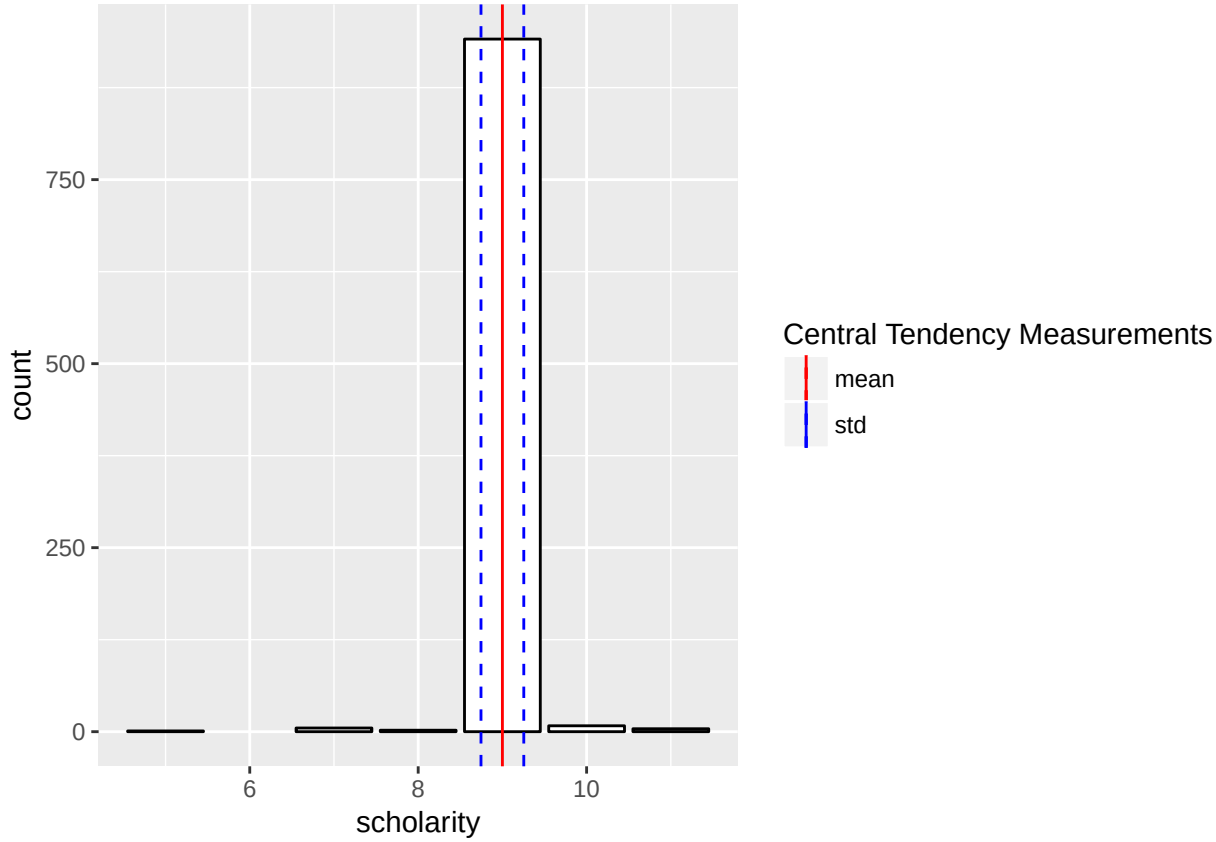
We have seven variables in this dataset: contract hours, age, gender, scholarity, average salary, minimum per hour, and employment time.

```
load(file="data/economist.Rdata")
summary(economist)
```

```
##    Scholarity contract_hours       age          avg_salary
##  Min.   : 5   Min.   : 8.00   Min.   :18.00   Min.   :  323.7
##  1st Qu.: 9   1st Qu.:40.00   1st Qu.:28.00   1st Qu.: 2000.0
##  Median : 9   Median :44.00   Median :33.00   Median : 3057.0
##  Mean   : 9   Mean   :41.75   Mean   :36.19   Mean   : 4418.7
##  3rd Qu.: 9   3rd Qu.:44.00   3rd Qu.:44.00   3rd Qu.: 5061.8
##  Max.   :11   Max.   :44.00   Max.   :69.00   Max.   :32180.3
##    min_salary         gender      employment_time
##  Min.   : 0.580   Min.   :1.000   Min.   :  0.40
##  1st Qu.: 3.660   1st Qu.:1.000   1st Qu.: 15.70
##  Median : 5.600   Median :1.000   Median : 40.90
##  Mean   : 8.109   Mean   :1.461   Mean   : 89.17
##  3rd Qu.: 9.280   3rd Qu.:2.000   3rd Qu.:101.90
##  Max.   :59.120   Max.   :2.000   Max.   :477.30
```

- Scholarity:

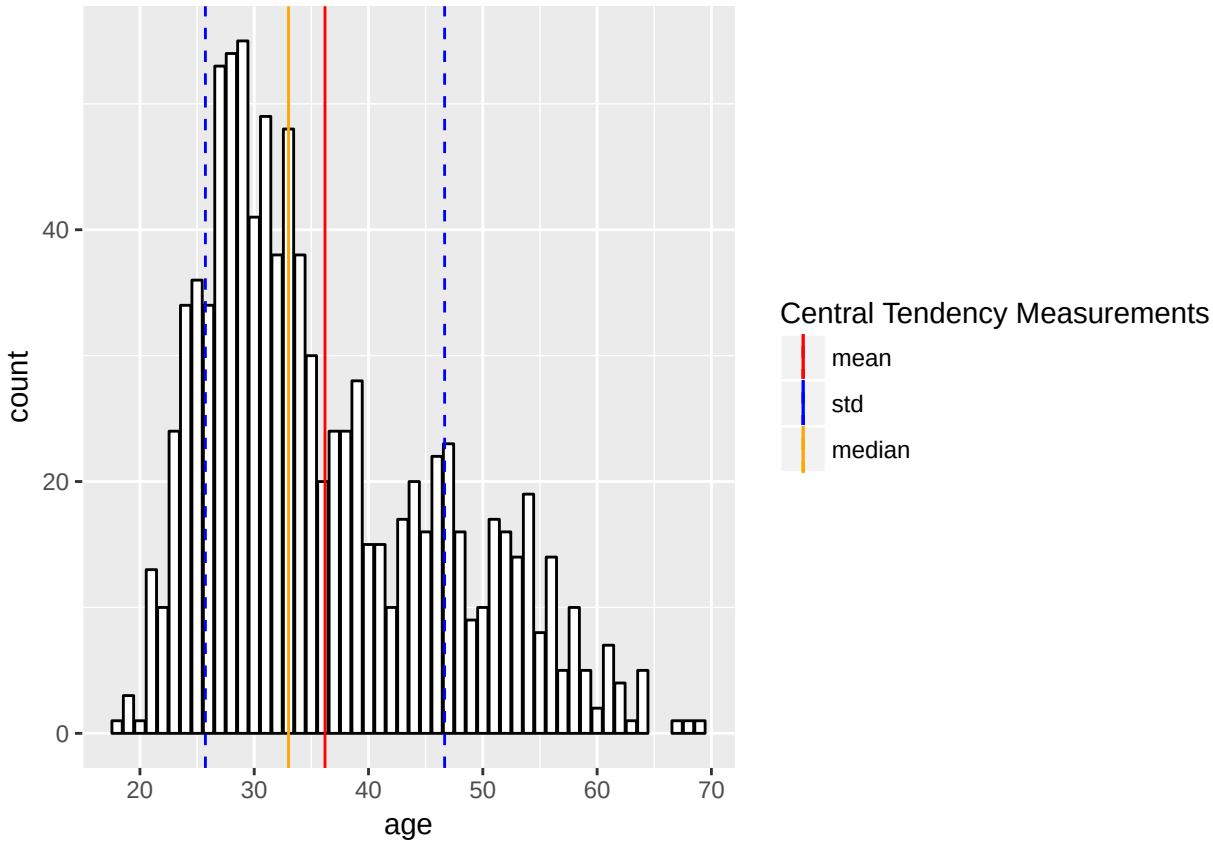
```
library(ggplot2)
meanE <- mean(economist$Scholarity)
std <- sd(economist$Scholarity)
ggplot(data = economist, aes(economist$Scholarity) ) + geom_bar(fill="white", colour = "black") + labs(
```



Looking at the difference of frequency between the different scholarity levels in the dataset, I believe is going to be difficult to infer things based on the level of scholarity, since most of the dataset falls in the category 9.

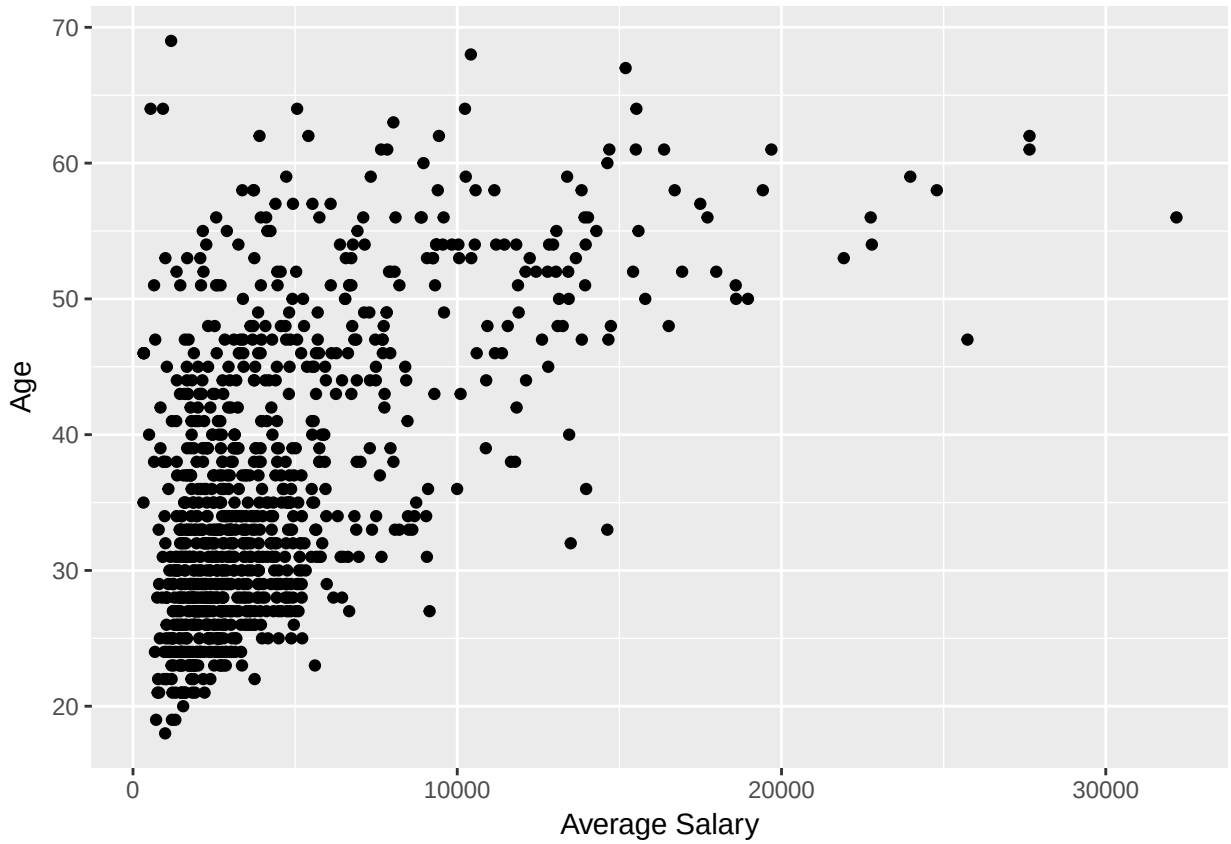
- Age:

```
meanE <- mean(economist$age)
std <- sd(economist$age)
ggplot(data = economist, aes(economist$age)) + geom_bar(fill="white", colour = "black") + labs(x= "age"
  scale_colour_manual(name = "Central Tendency Measurements", breaks = c("mean", "std","median"), value
```



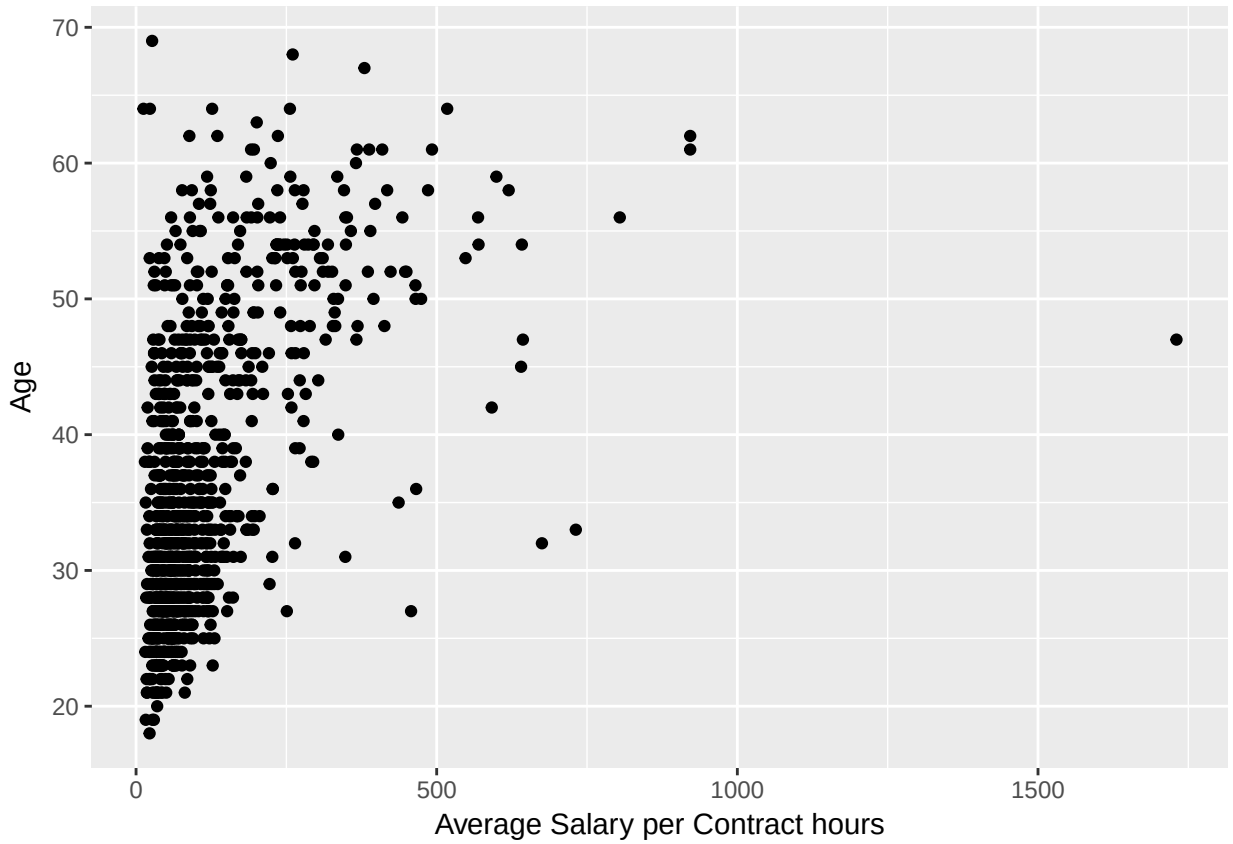
age x salary

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



However, here we are not considering the amount of hours that each economist is doing per month, and it can give us the false impression that someone young is receiving a lot more money, where in reality they could only be working more.

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


70
60
50
40
30
20
Age
0
500
1000
1500
Average Salary per Contract hours