U.S. Women's Labor-Force Participation

Ebrahim Naehimi, Mauricio Vazquez and Charles Bailly M2 MoSIG, January 2017

Our work is based on data about the women's labor force participation in the USA. It contains information about 753 married women, collected by the Panel Study of Income Dynamics. The dataset we concretely used was extracted from a paper called *The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions* [T. Mroz, 1987]. The dataset is available on this page: https://vincentarelbundock.github.io/Rdatasets/csv/car/Mroz.csv

Dataset

The dataset we chose contains a lot of information about these women and their husband: age, wage, education level, children, experience, husband's wage, and more. When we started to work on it, we had several basic hypothesis in mind:

Hypothesis 1: Women with working experience have higher wages than those who don't have.

Hypothesis 2: Women earn less money than their husband

Hypothesis 3: Women work less if they have children

In addition to checking these intuitions, we were also curious of what we could find by exploring the dataset.

First steps

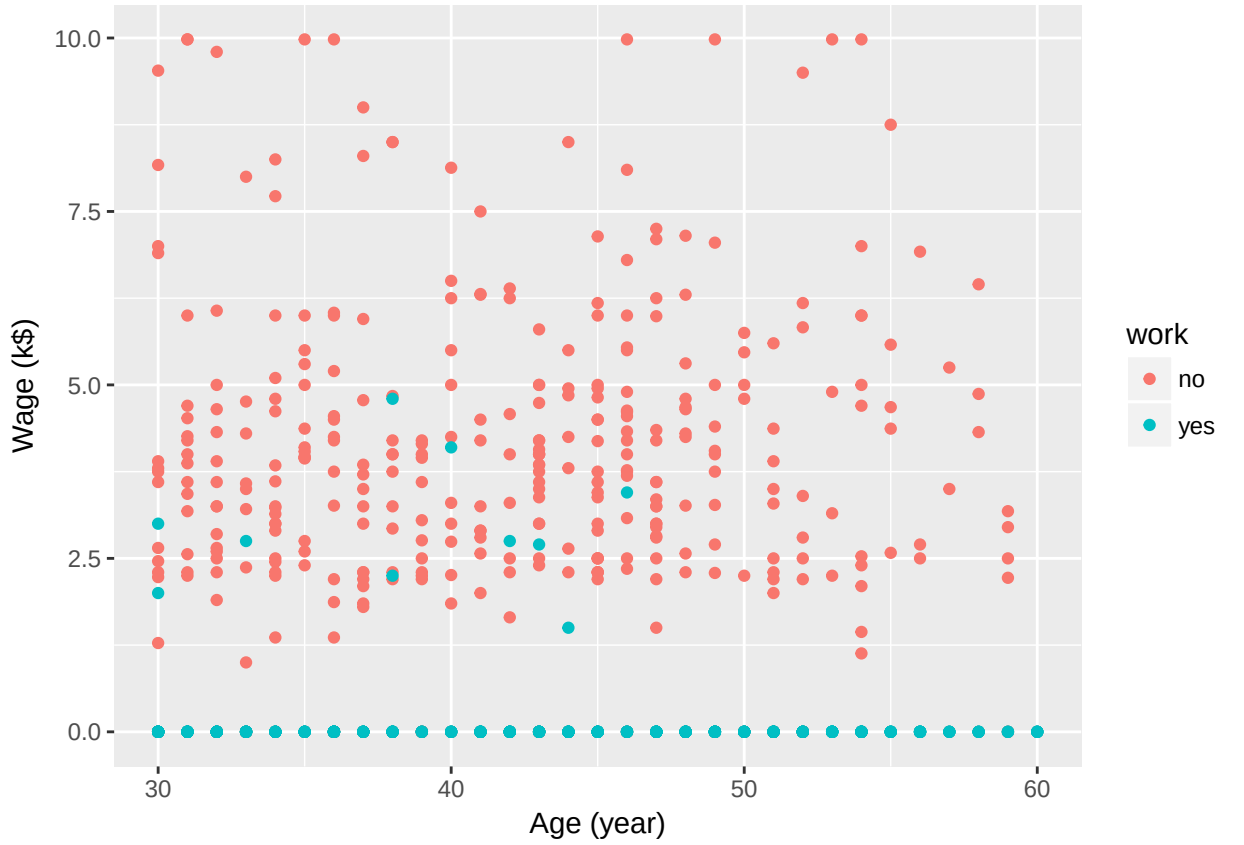
Since there is a lot of information in the dataset (18 variables), We first plotted a few basic statistics about the main columns:

```
library(ggplot2)
library(plyr)
library(reshape2)
my_data = read.csv("./Data/Mroz.csv")

# Not all columns are inreresting to sum up (too much otherwise)
# Here, we focus on columns about women
summary(my_data[c(3, 4, 5, 6, 9, 17, 19)])
```

```
##      hoursw           child6          child618          agew      
##  Min.   :   0.0   Min.   :0.0000   Min.   :0.000   Min.   :30.00  
##  1st Qu.:   0.0   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00  
##  Median : 288.0   Median :0.0000   Median :1.000   Median :43.00  
##  Mean   : 740.6   Mean   :0.2377   Mean   :1.353   Mean   :42.54  
##  3rd Qu.:1516.0   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:49.00  
##  Max.   :4950.0   Max.   :3.0000   Max.   :8.000   Max.   :60.00  
##      wagew        unemprate       experience   
##  Min.   :0.00   Min.   : 3.000   Min.   : 0.00  
##  1st Qu.:0.00   1st Qu.: 7.500   1st Qu.: 4.00  
##  Median :0.00   Median : 7.500   Median : 9.00  
##  Mean   :1.85   Mean   : 8.624   Mean   :10.63  
##  3rd Qu.:3.58   3rd Qu.:11.000   3rd Qu.:15.00  
##  Max.   :9.98   Max.   :14.000   Max.   :45.00
```

Then, we started creating plots to explore the dataset. However, we quickly found something strange thanks to the following one:

```
p0 <- ggplot(data=my_data, aes(x=agew, y=wagew, color=work)) +
  labs(x="Age (year)", y="Wage (k$)") +
  geom_point()
p0
```



According to the colour mapping, it seems that most of women who were not working where earning much more money than those with a job! In fact, except for a few women, working leads to a null wage... This was unexpected for sure, so we went back to the raw data to check if anything wrong happened (a bad manipulation, for example). But no, the raw data was really this one, we didn't made an error so far. We quickly discussed about this phenomenon, and also linked it with the number of hours worked. The result was the same: nearly all women without job had a lot of hours worked, and those with a professional activity didn't (0 most of the time). Therefore, we finally decided to consider it as an error in the raw data, and worked on a cleaned version of it.

```
data_clean = read.csv("./Data/Mroz_cleaned.csv")
```

Experience and wage

Our first hypothesis was that experienced women earn more money than those without experience. We began by creating an histogram to have an idea of the experience of women:

```
# Stats
median_exp <- median(data_clean$experience)
mean_exp <- mean(data_clean$experience)
sd_exp <- sd(data_clean$experience)
```

```r
#Display stats
median_exp
```

```
## [1] 9
```

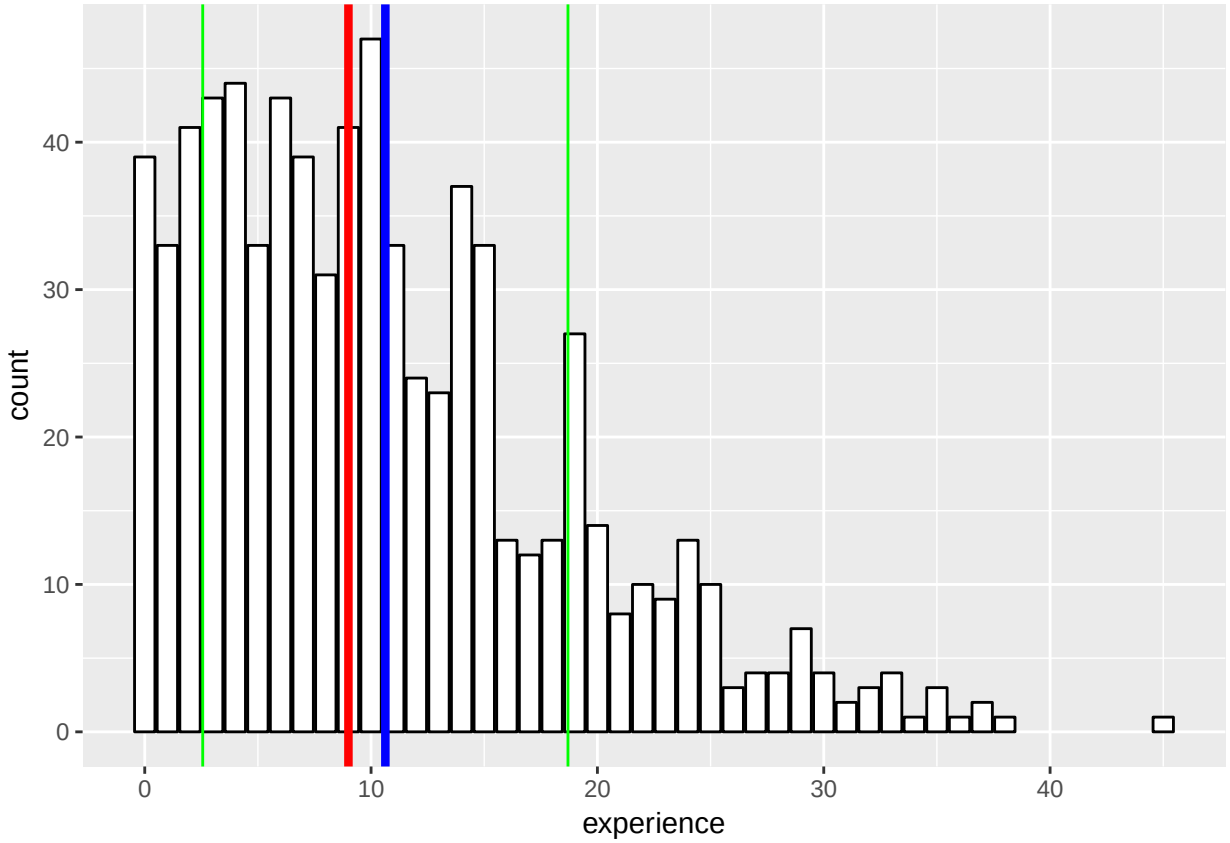
```r
mean_exp
```

```
## [1] 10.63081
```

```r
sd_exp
```

```
## [1] 8.06913
```

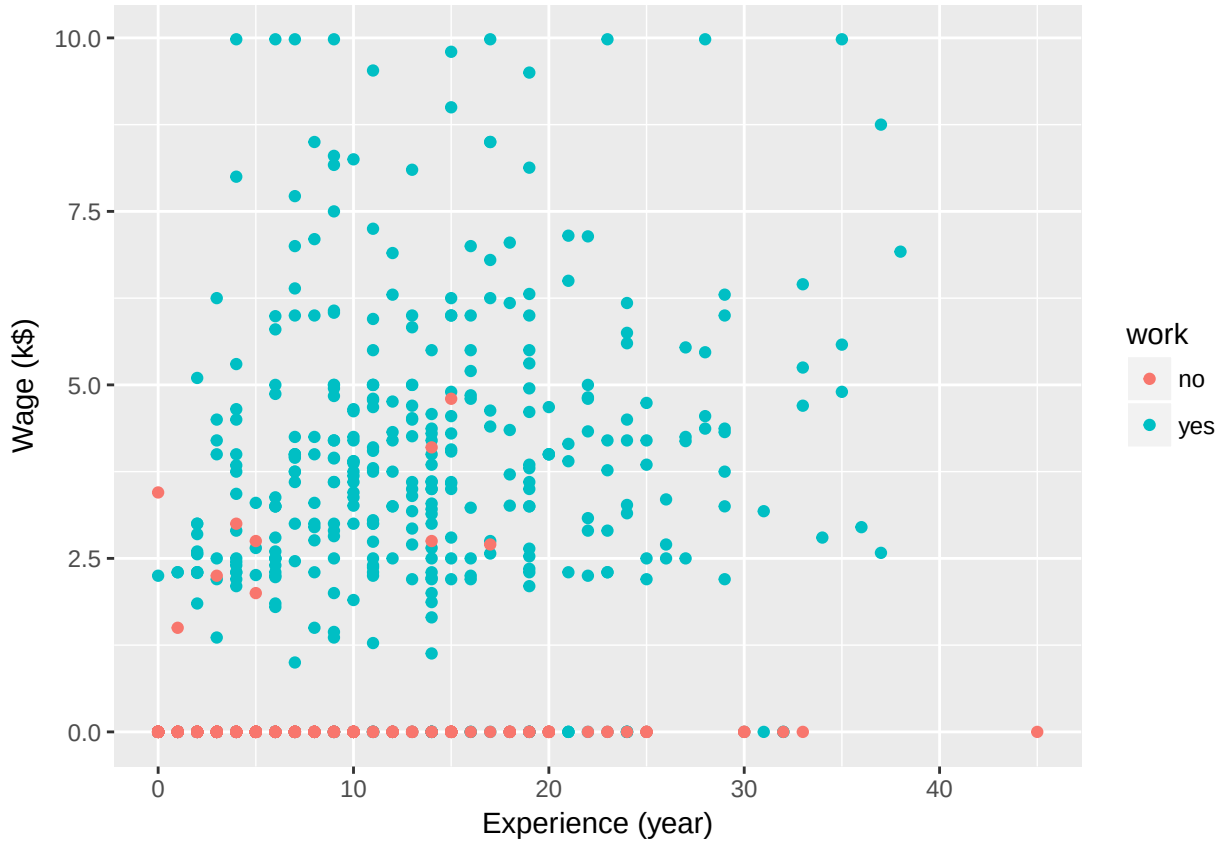
```r
#Plot
p1 <- ggplot(data=data_clean, aes(experience)) +
  geom_bar(fill="white", colour = "black") +
  geom_vline(xintercept = median_exp, color="red", size = 1.5) +
  geom_vline(xintercept = mean_exp, colour = "blue", size = 1.5) +
  geom_vline(xintercept = mean_exp-sd_exp, colour="green") +
  geom_vline(xintercept = mean_exp+sd_exp, colour="green")

p1
```



In average, women had 10.6 years of working experience, but the standart deviation is quite large. We are not expert on the field, so we can't affirm that it matches with the historical reality of the USA in 1978.

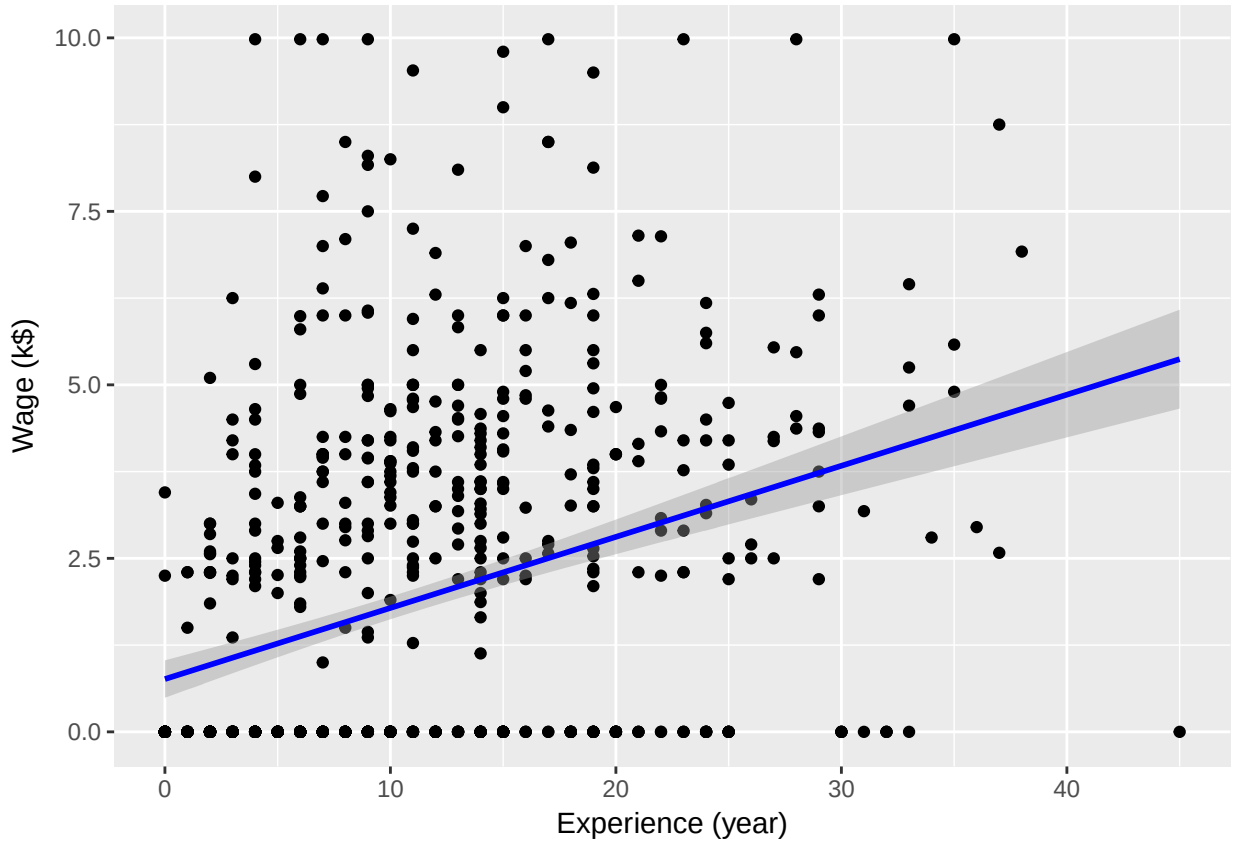
Then, we did a very basic graph to see if there was an obvious relationship between experience and wage:

```
p2 <- ggplot(data=data_clean, aes(experience, wagew, color=work)) +
  geom_point() +
  labs(x="Experience (year)", y="Wage (k$)")

p2
```



We already had the feeling that we could not fit very well a linear model on it:

```
p3 <- ggplot(data=data_clean, aes(experience, wagew)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  labs(x="Experience (year)", y="Wage (k$)")

p3
```



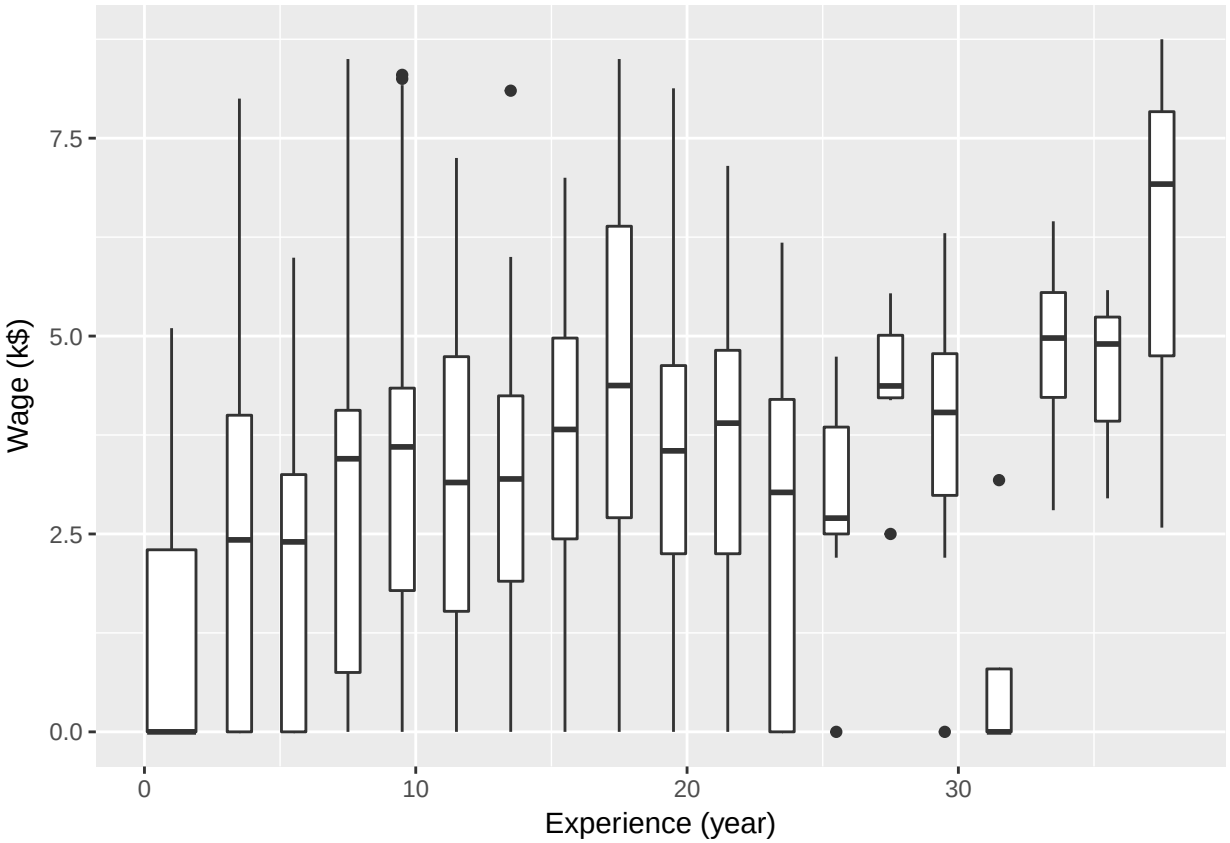

```
# Check exact correlation between the two
cor(data_clean$experience, data_clean$wagew)
```

```
## [1] 0.3415569
```

As we thought, the correlation between experience and wage is positive but quite low.

We tried another approach with boxplots:

```
# Women who are working, without outliers(huge wages)
data_without_hW <- subset(data_clean, data_clean$work=="yes" & data_clean$wagew<9 )

p4 <- ggplot(data=data_without_hW, aes(x=experience, y=wagew, group=cut_interval(x = experience, length
  geom_boxplot() +
  labs(x="Experience (year)", y="Wage (k$)")

p4
```



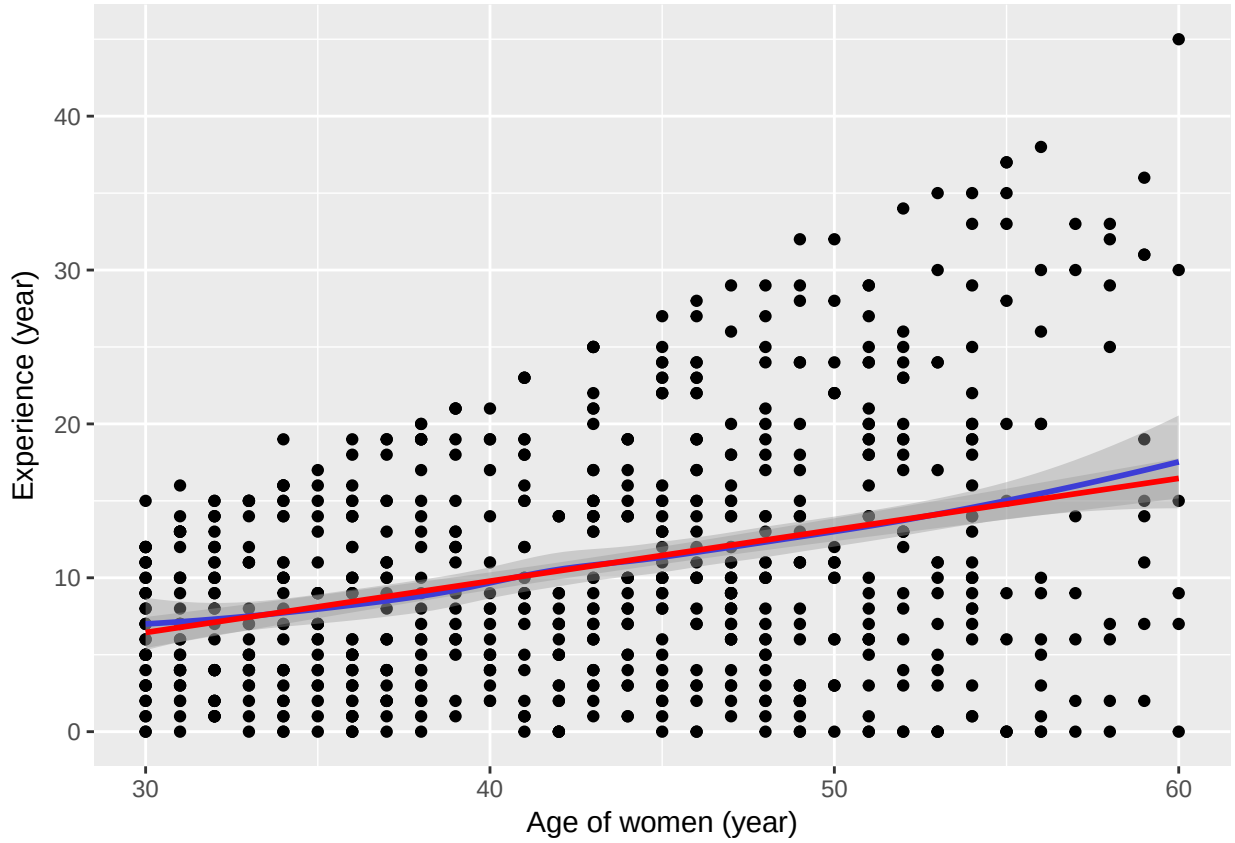
At first sight, it seems that we can conclude that there are significant differentes between women who have very little experience and those who have a lot. However, the number of women with more than 30 years of experience is really limited. Therefore,it is hard to find any firm conclusion about it.

Experience vs age

The previous work was a bit unconclusive, so we tried to find some data where we hopefully could find a clearer result. We thought that experience should be somehow related to the age, and started working on it.

```
working_w <- subset(data_clean, data_clean$work=="yes")

p5 <- ggplot(data=data_clean, aes(agew, experience)) +
  geom_point() +
  geom_smooth(color = "blue") + #loess method
  geom_smooth(method = "lm", color="red") +
  labs(x="Age of women (year)", y="Experience (year)")

p5
```

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




```
cor(data_clean$agew, data_clean$experience)
```

```
## [1] 0.3340159
```

```
cor(working_w$agew, working_w$experience)
```

```
## [1] 0.4836462
```

We expected a much greater correlation between age and experience. If we consider only the subset of working women, it is slightly better, but once more not enough to conclude positively.

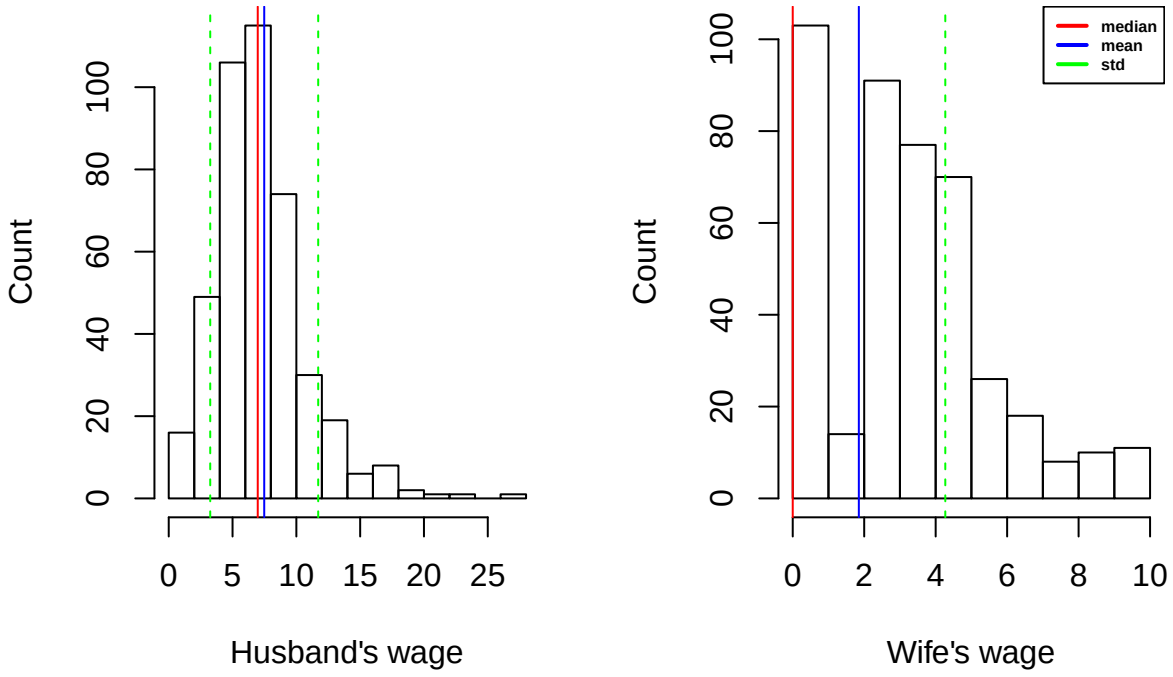
Comparing wives and husbands

Our third focus was about the wage of women and husbands.

First, we tried to check if husbands earn more money or not:

```
par(mfrow=c(1,2))

hist(working_w$wageh,xlab="Husband's wage", ylab="Count", main="")
abline(v = medianh, col = "red")
abline(v = meanh, col = "blue")
abline(v = meanh+stdh, col = "green", lty = 2)
abline(v = meanh-stdh, col = "green", lty = 2)

hist(working_w$wagew,xlab="Wife's wage", ylab="Count",main="")
abline(v = medianw, col = "red")
```

```
abline(v = meanw, col = "blue")
abline(v = meanw+stdw, col = "green", lty = 2)
abline(v = meanw-stdw, col = "green", lty = 2)
legend("topright",c("median","mean","std"),col=c("red","blue","green"),lwd=2,text.font=2,cex = 0.5)
```



There are several things to say here.

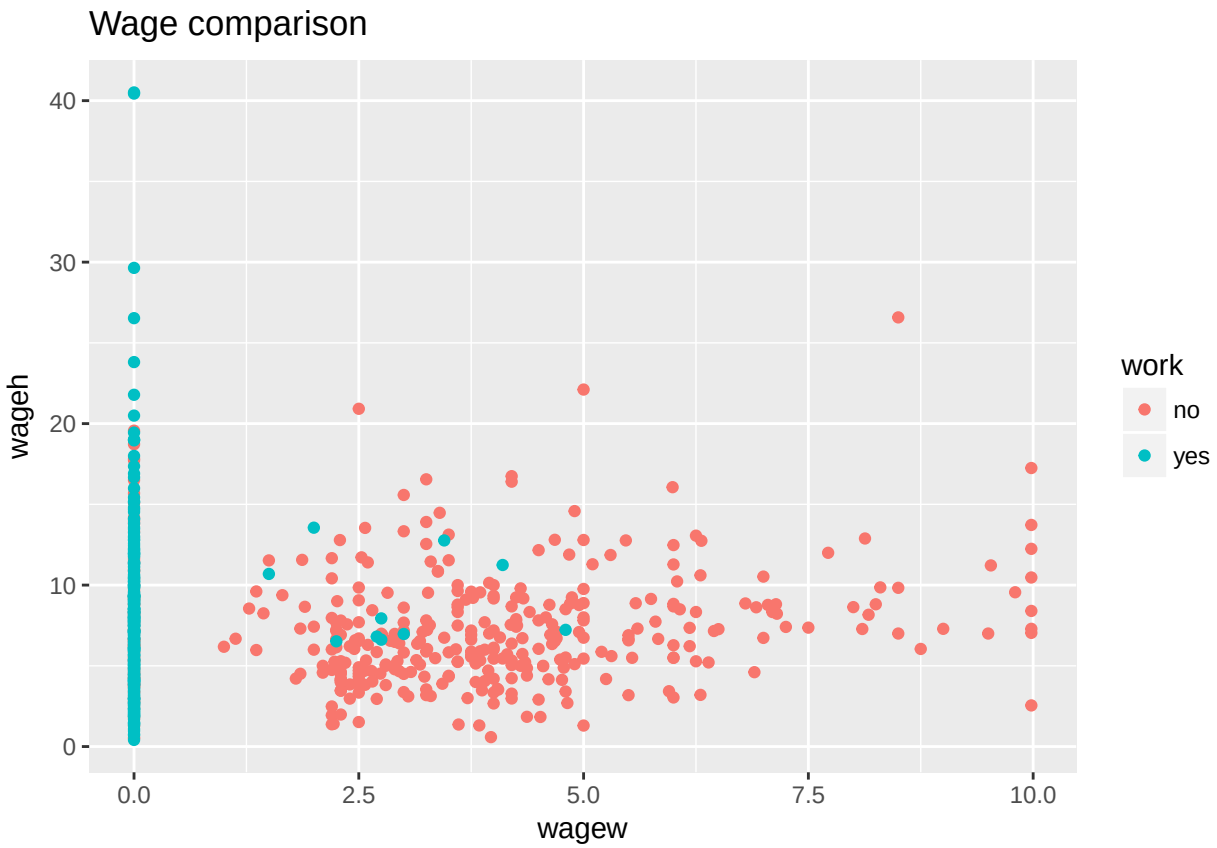
First, the x_axis of the two graph is not the same.

Then, the data for husbands seems to correspond to a normal distribution, while for women we don't really know.

We can see than a lot of husbands earn more money than there wifes indeed, which was one of our hypothesis.

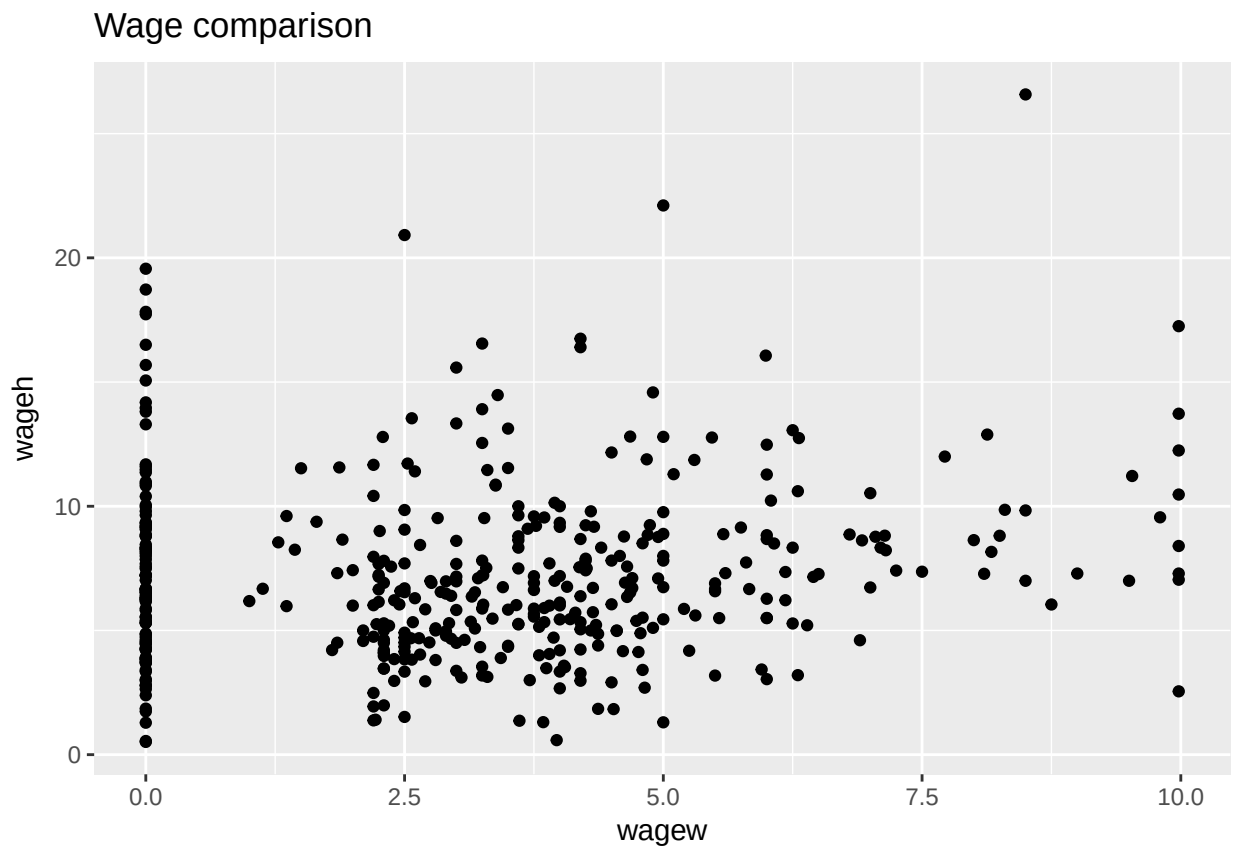
Then, we tried to find a precise relation between the two, plotting both wagew and wageh together:

```
qplot(data=my_data ,x=wagew, y=wageh,color=work,main="Wage comparison")
```



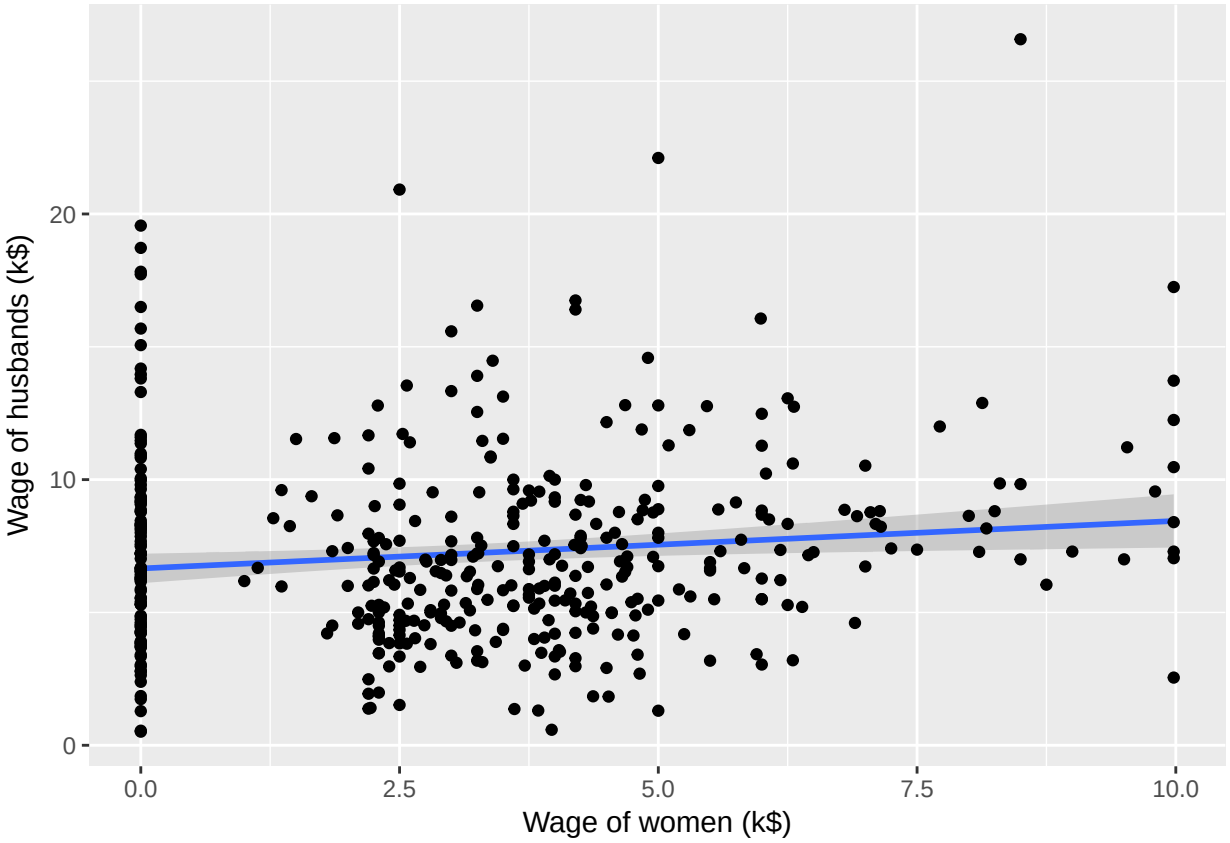
It's observed that there's nothing meaningful. Regarding substitution of "work" column values, we remove values having work=yes to obtain valide wages for people who are working

```
qplot(data=working_w,x=wagew, y=wageh,main="Wage comparison")
```



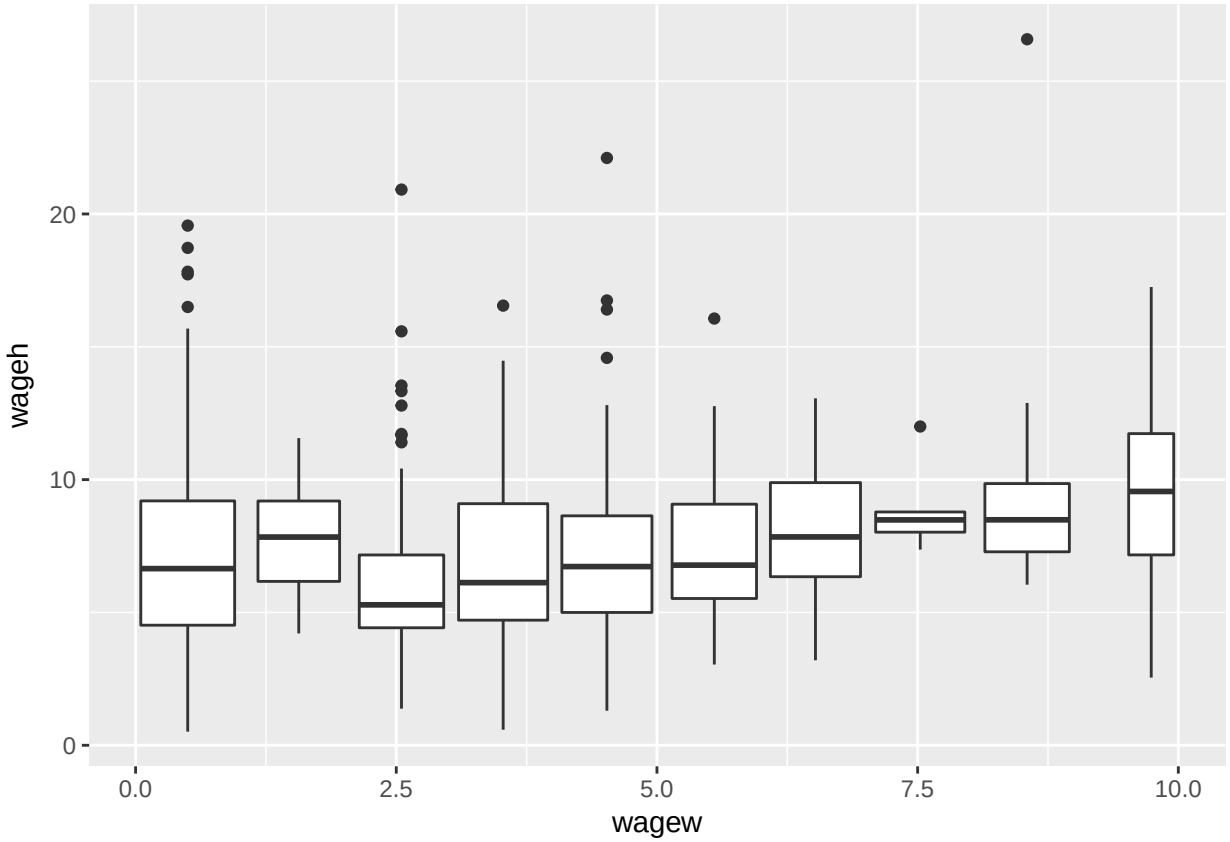
We tried to fit a linear model, and this is what we obtained:

```
ggplot(data=working_w ,aes(x=wagew, y=wageh)) +
  geom_smooth(method="lm") +
  labs(x="Wage of women (k$)", y="Wage of husbands (k$)") +
  geom_point()
```



We also tried to group data regarding wife's wage and plotted related box-plot:

```
ggplot(data=working_w, aes(x=wagew, y=wageh , group=cut_interval(x=wagew, length=1))) +
  geom_boxplot()
```

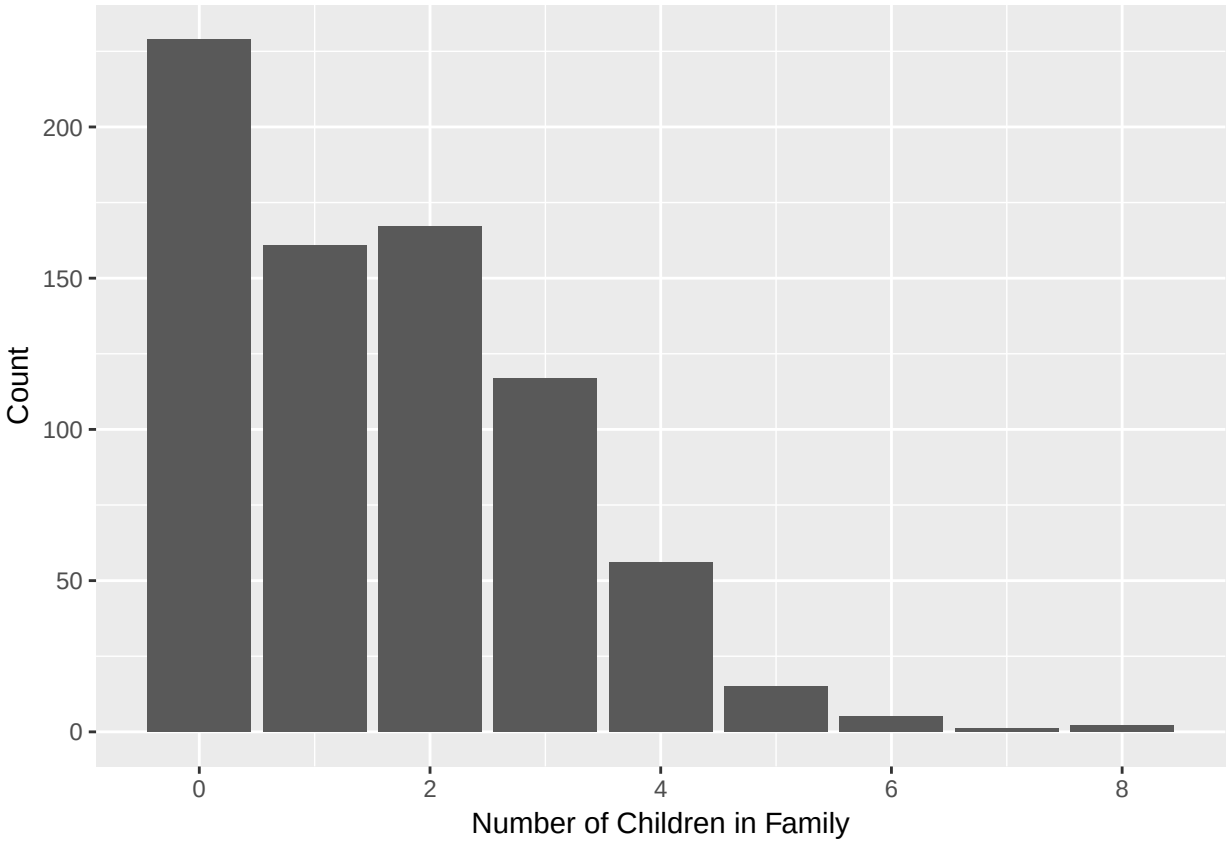


Neither boxplots nor linear regression allow us to conclude anything about a precise relation between the wage of women and the one of their husband.

Comparing work and children

We were curious to see whether or not the amount of work done by the husband and wife have any effect on the children that the couple have. But first, we wanted to see what the total amount of children in a household was. In our data, we have the information regarding children aged from 0 to 5 years old, and 6 to 18 years old. We grouped them together to get this information.

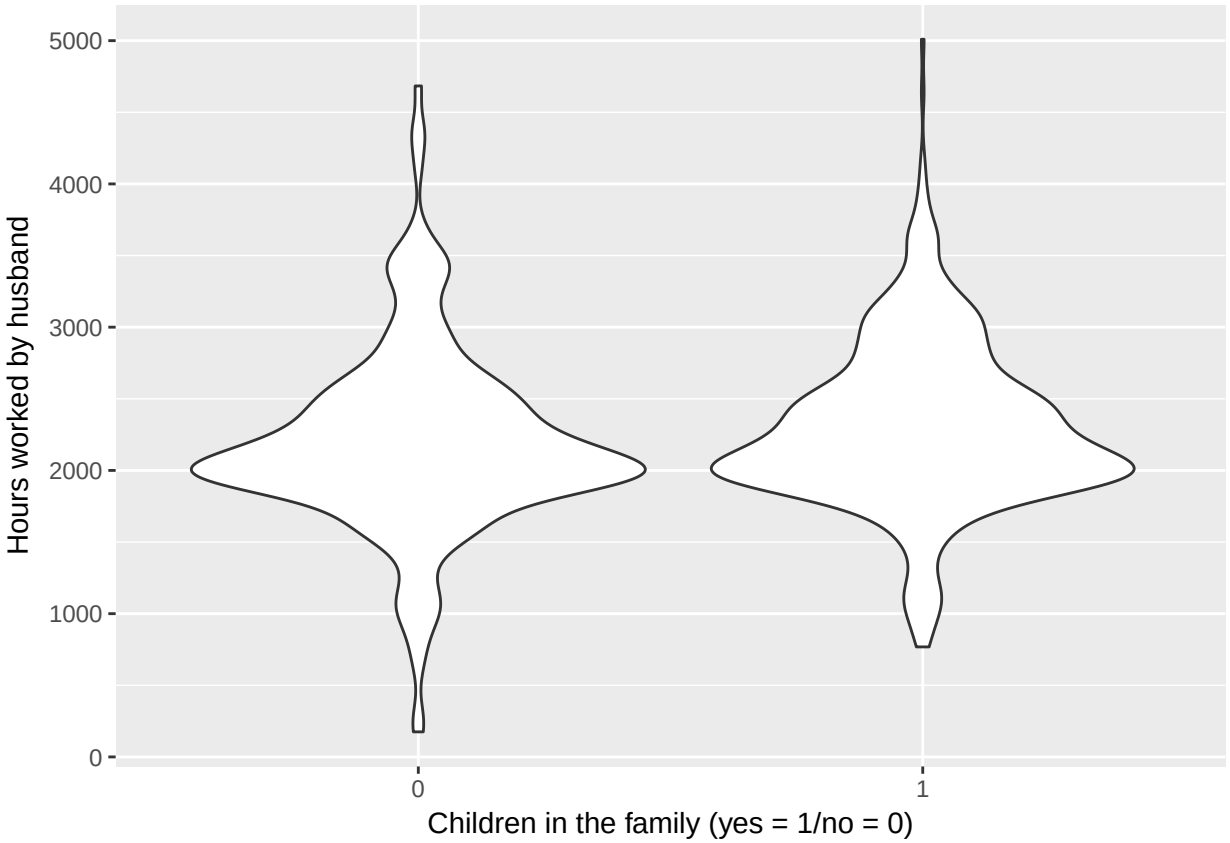
```
ggplot(my_data, aes(child6+child618)) + geom_bar() + xlab("Number of Children in Family") +ylab("Count"
```



In the previous figure, we can see that the majority of the families do not have any children. Based on our initial quest to find a relationship between work and children, it would be very interesting to see whether or not this big group of families do not have children due to the work hours they are doing.

We continued by plotting the amount of work hours done by the husband, the wife, and both, to see if one, both, or none of their work hours have any effect on the amount of children in the household.

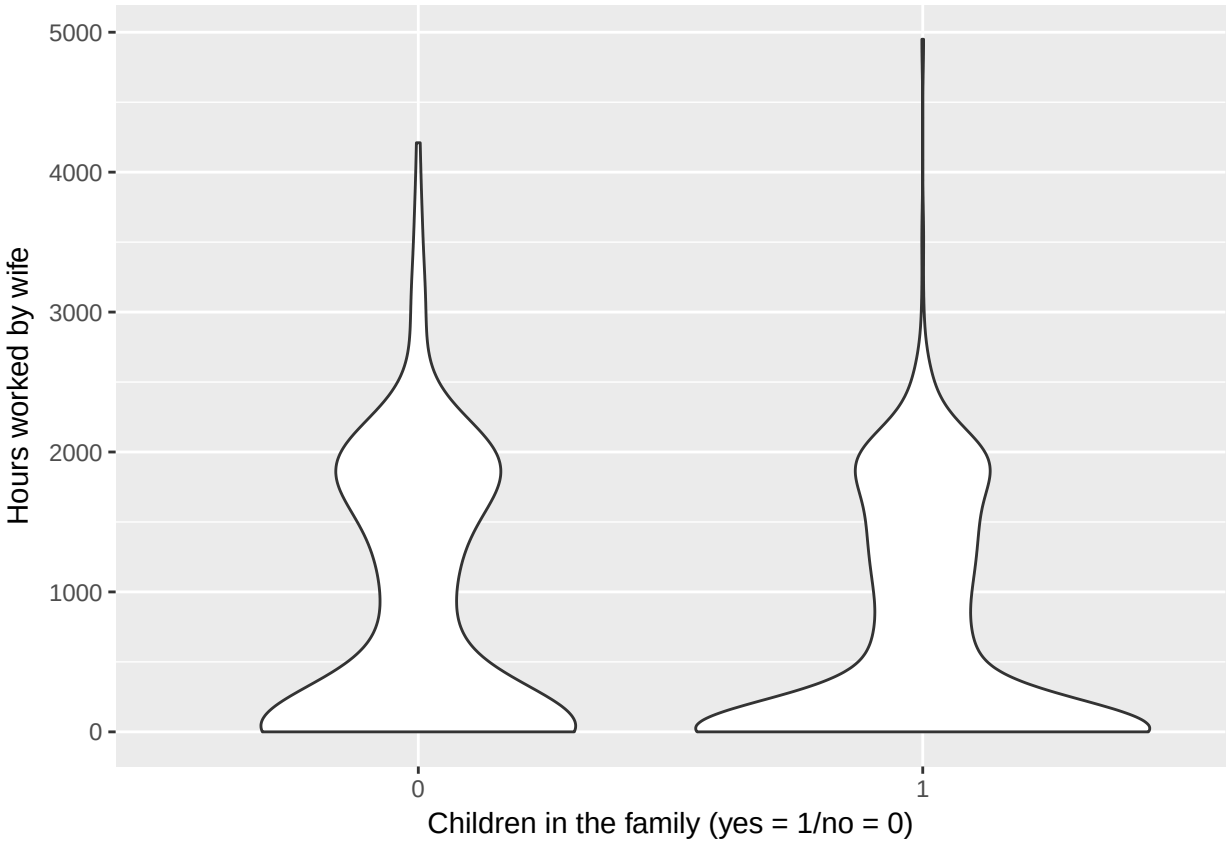
```
ggplot(my_data, aes( factor(ifelse((child6+child618) > 0, 1, 0)), hoursh)) + geom_violin(scale="area") +
```




```
cor(my_data$hoursw,ifelse((my_data$child6+my_data$child618) > 0, 1, 0))
```

```
## [1] -0.0966394
```

```
ggplot(my_data, aes( factor(ifelse((child6+child618) > 0, 1, 0)), hoursw)) + geom_violin(scale="area") +
```




```
cor(my_data$hoursh,ifelse((my_data$child6+my_data$child618) > 0, 1, 0))
```

```
## [1] 0.08244192
```

```
ggplot(my_data, aes( factor(ifelse((child6+child618) > 0, 1, 0)), hoursw+hoursh)) + geom_violin(scale="a
```


```
cor((my_data$hoursh + my_data$hoursw),ifelse((my_data$child6+my_data$child618) > 0, 1, 0))
```

```
## [1] -0.03416976
```

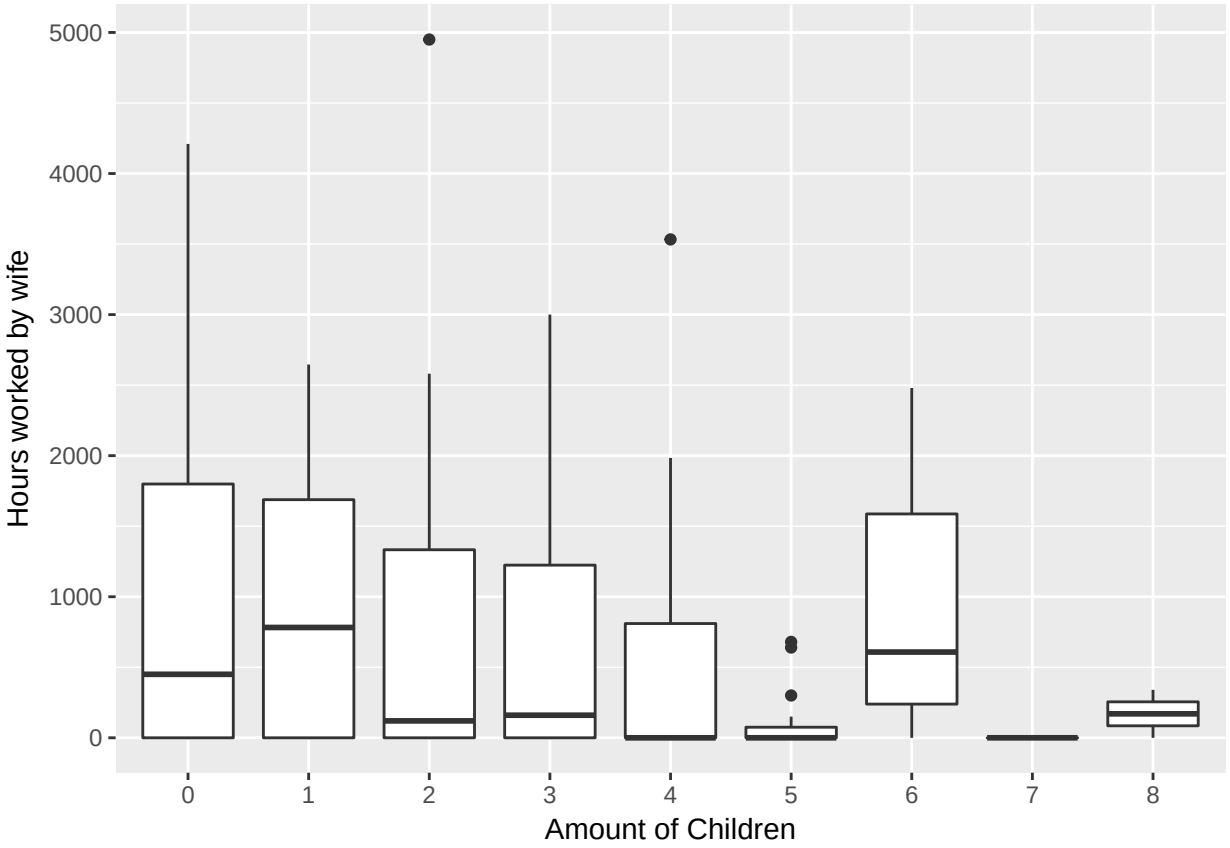
In the first graph of hours worked by the husband, we failed to see any relationship between the amount of work, and whether or not the family has children. Although we did not find a way to compare the violin plots from the husband graph with something similar to a confidence interval, we find it interesting to see that the graph of those men with children is slightly more stretched towards the top, meaning that men with children would have a slightly higher workload.

The second graph represents the hours worked by the women, and whether or not they have children. We can immediately see that both plots for the women (with and without children) are significantly lower than those for their husbands. We interpreted this as women having either no work, or less workload, but just based on this we cannot certainly say that work results in no children. As with the previous plot, we did not find a way to make a confidence interval comparison between both women with children, and those without. However, we found it interesting that contrary to men, and as we expected, women without children work more than those with children.

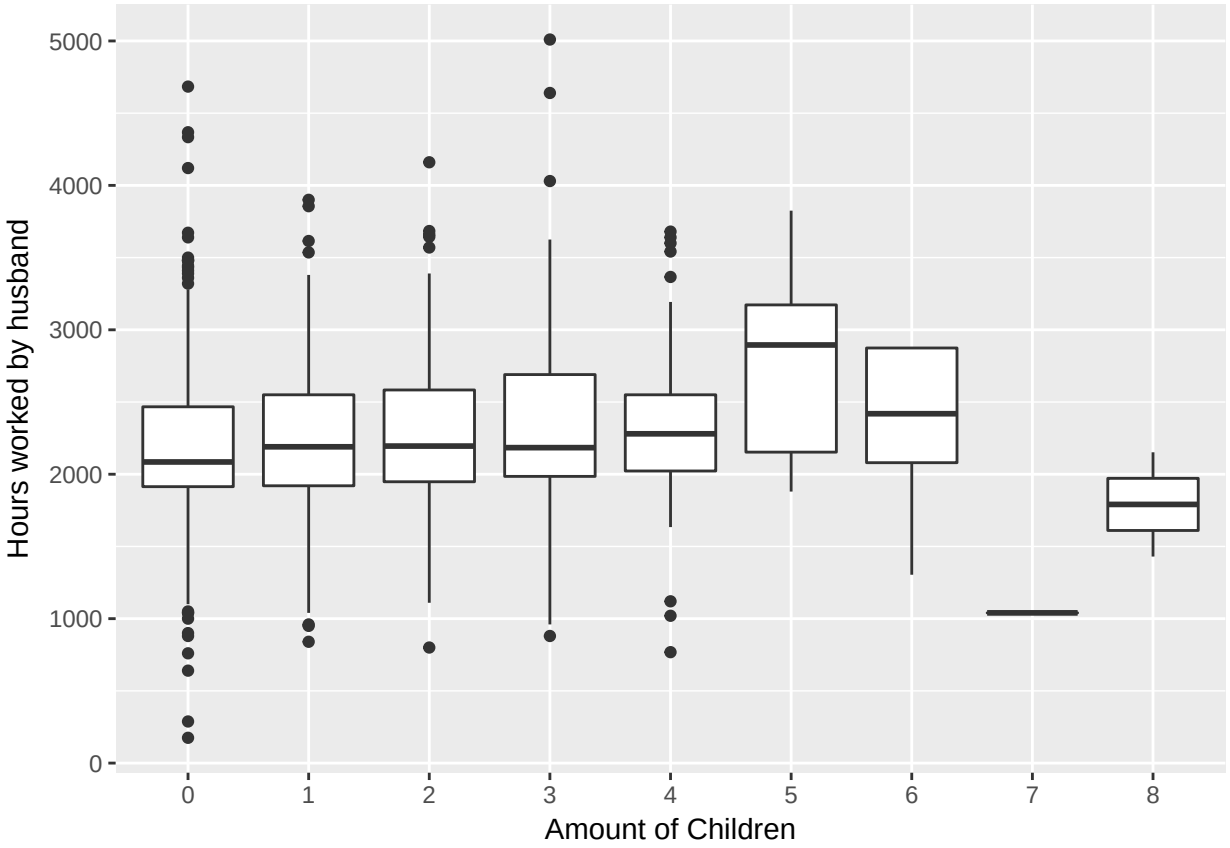
We finally added a third graph comparing the amount of children in a household against the amount of hours worked by both the husband and the wife. We can observe that there is a very similar amount of activity done by both groups, probably because of the inverse work responsability done by husband and wife, where husbands work more when there are children, and women work more when there are no children.

Finally, we were interested in observing whether the amount of work has an impact on the amount of children in a family:

```
ggplot(my_data, aes( factor(child6+child618), hoursw)) + geom_boxplot() + xlab("Amount of Children") + y
```




```
ggplot(my_data, aes( factor(child6+child618), hoursh)) + geom_boxplot() + xlab("Amount of Children") + y
```




```
cor(my_data$hoursw, my_data$child6 + my_data$child618 )
```

```
## [1] -0.1615735
```

```
cor(my_data$hoursh, my_data$child6 + my_data$child618 )
```

```
## [1] 0.0985254
```

In the first graph, were we compare the amount of children against the hours worked by the wife. We observe that, although some outliers exist, that as the amount of children in the family increases, the hours worked by women decrease. This is not true however for the case of 6 children. On the first graph, we compare the amount of children against the hours worked by the husband. We can observe that there is a slight increase in workload as the amount of children increase.

During this analysis, we saw how women with children tend to work less than those without, and alternatively men with children work more than those without. We then show how this same pattern can be seen in the amount of children in a family, more so on women than on men. To answer our inital question of whether the amount of work has an impact on the amount of children in a family, we converge to the idea that the effect is the other way around, where children have an effect on the amount of workload the parents have.

Conclusion

We explored the dataset trying to confirm our hypothesis, but it was not so easy. We were not able to prove that Women with working experience have higher wages than those who don't have, which was our first hypothesis. However, we saw that women are less payed than their husbands, and that children impact the number of working hours of their parents.

References

[1] T. A. Mroz (1987) *The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions.* Econometrica 55, 765-799