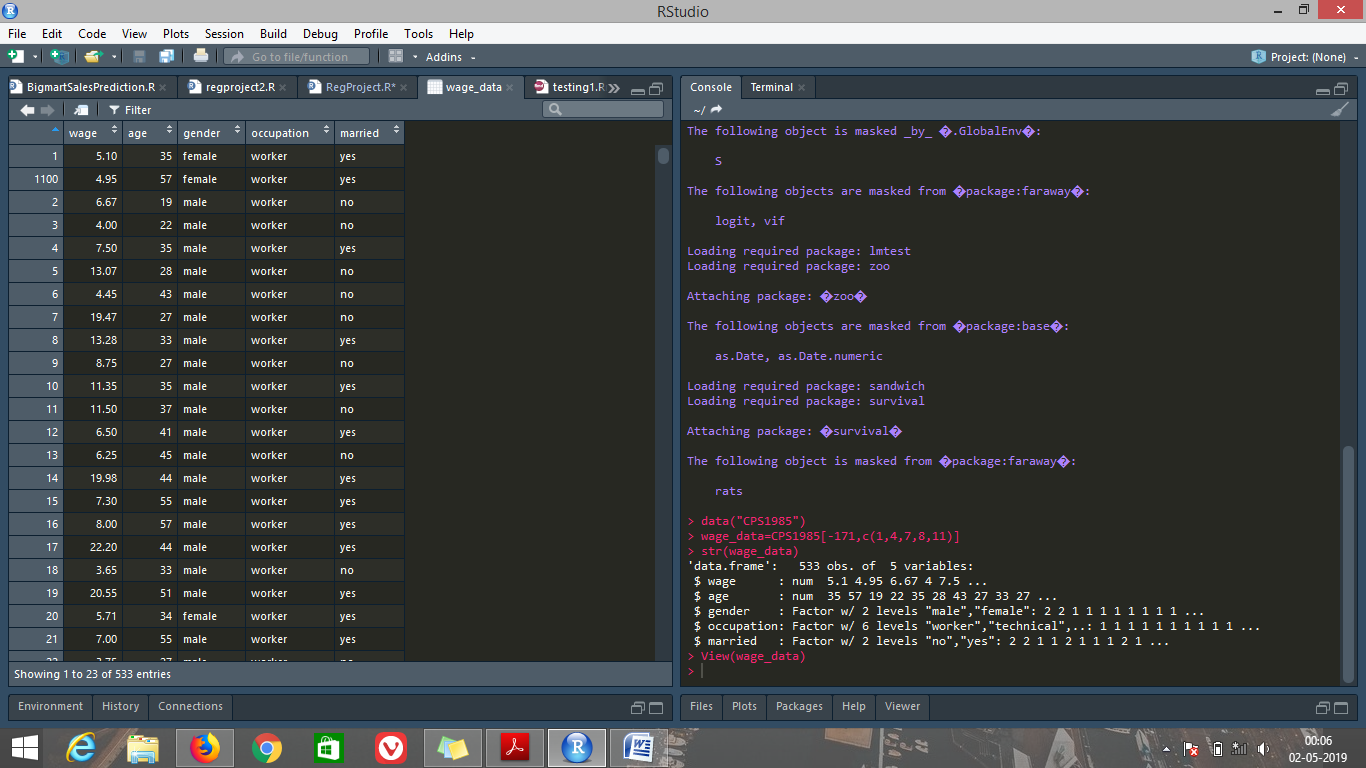
Exploratory Data Analysis

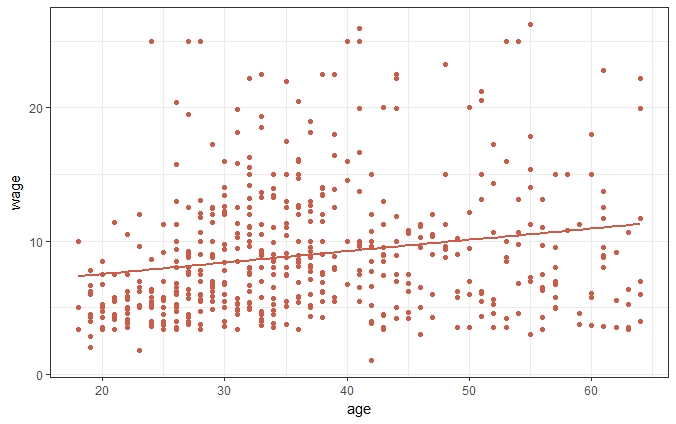
In this post, we consider a simple dataset and consider analysing it statistically. A glimpse of the dataset is as follows. It consists of information on wage, age, gender, occupation and marital status of 533 individuals, obtained from some suitable survey.



We want to understand how wage varies with respect to age and other variables. The focus is not to model ‘wage’ on others but to simply understand its dependence on other variables using graphs and simple linear models. This kind of analysis, prior to building a model formally, helps tremendously.

We start with ‘age’ and ‘wage’.

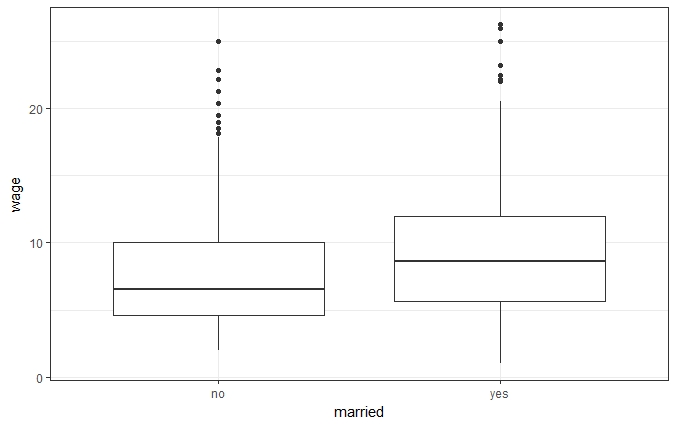
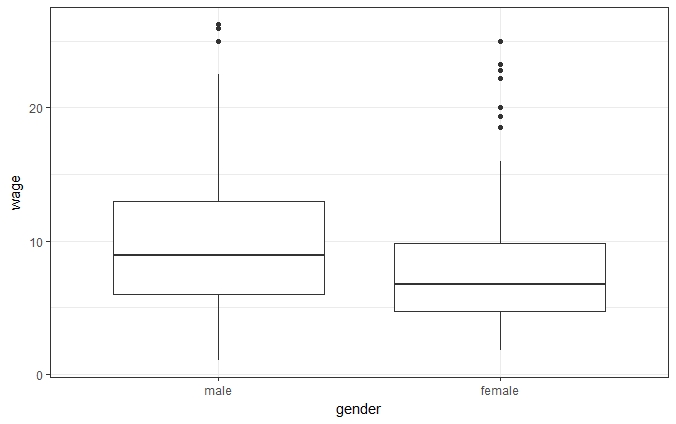
A scatterplot is the best way to understand the relationship between two numeric variables.



We also added the linear regression line of wage on age. This gives the idea that with increasing age, wage seems to increase, i.e. as a person grows older, he/she seems to earn more.

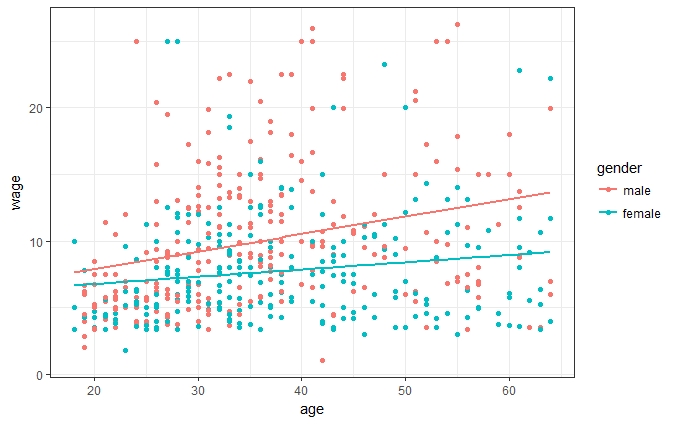
Next we also bring in ‘gender’ & ‘married’, to understand whether the relationship remains same with gender and marital status.

Initially, boxplots are considered.

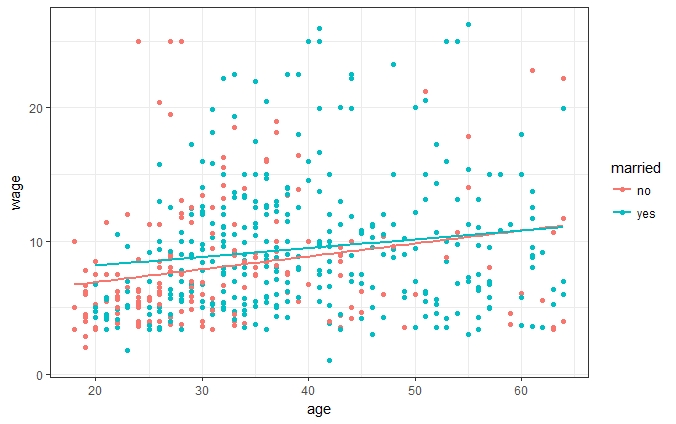


1. The median income for males is higher than that of females. Similarly, married people are earning more than unmarried ones.
2. Variation in wage is quite similar in both the groups in both the graphs.

Along with ‘age’,

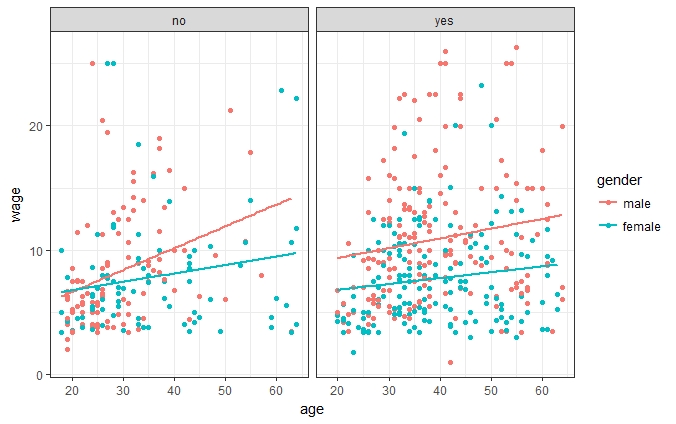


A male earns more than a female of the same age, and this gap widens as they grow older.



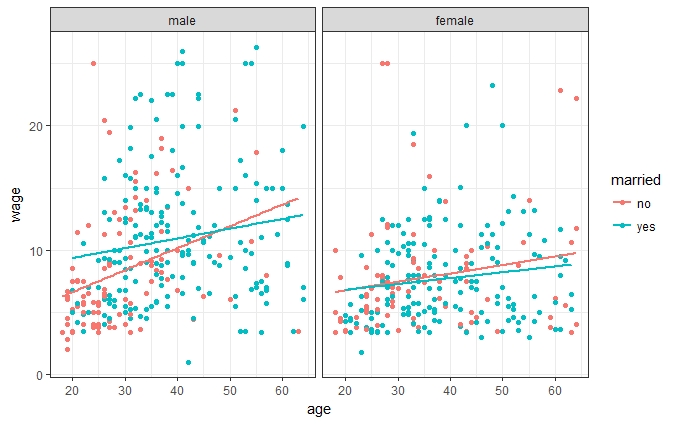
A married person, below 45 years of age, seems to earn more than an unmarried one, of the same age. Above 45 years of age, the gap is not visible.

From above, it seems that gender of a person has a strong effect on his/her wage. Both these graphs doesn’t consider the effect of the third variable, adding which might give us a better understanding.



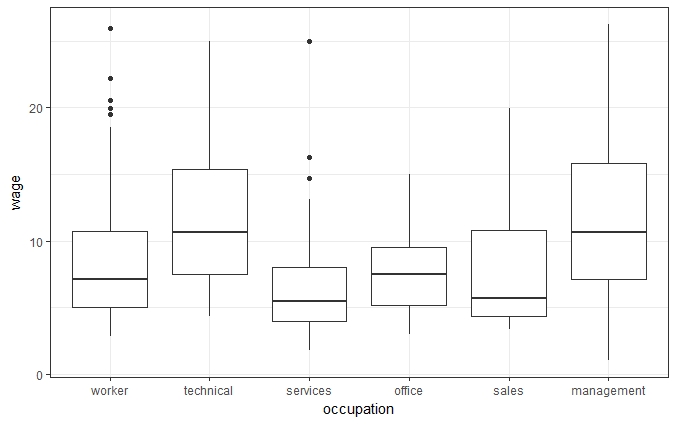
A married man seems to earn more than a married woman of same age, and this gap doesn’t change with age. An unmarried man, who is above 30 years of age, also seems to earn more than an unmarried woman. However, unmarried men and women who are less than 30 years of age, seem to earn the same.

Also, it seems that married women are earning lesser compared to unmarried ones and this relationship is opposite for males.



1. For males, at younger age, married ones are earning more, but as they grew older, only a few males remain unmarried, and they seem to work in high paying jobs, thus for unmarried ones, wage increases sharply and for married ones are earning more but that is stable with age.
2. For females on the other hand, this gap doesn’t seem to be prominent. The difference is only visible for older females.
3. Although we were expecting a much wider gap for females, this plot doesn’t suggest the same. Thus it requires further investigation.

Now we explore wages in different occupations.



We observe that ‘occupation’ has six categories, some gives higher wages than others. In order to reduce the no. of levels consider a linear model of ‘wage’ on ‘occupation’.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.4265 0.3575 23.568 < 2e-16 \*\*\*

occupationtechnical 3.5210 0.5637 6.246 8.67e-10 \*\*\*

occupationservices -1.8890 0.6067 -3.114 0.00195 \*\*

occupationoffice -1.0039 0.5774 -1.739 0.08269 .

occupationsales -0.8338 0.8078 -1.032 0.30246

occupationmanagement 3.6887 0.7051 5.232 2.43e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.466 on 527 degrees of freedom

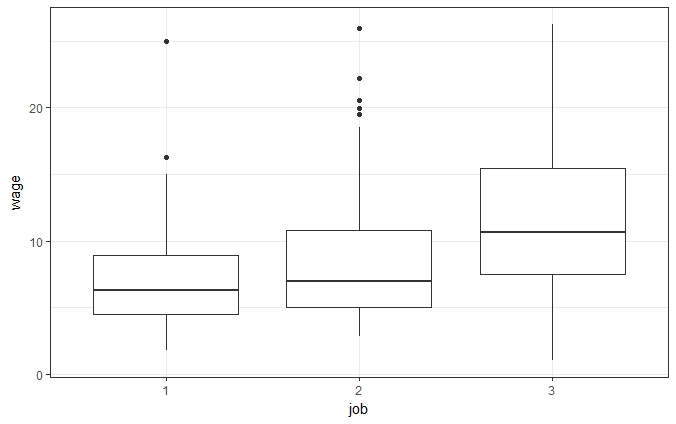
Multiple R-squared: 0.18, Adjusted R-squared: 0.1722

F-statistic: 23.13 on 5 and 527 DF, p-value: < 2.2e-16

Thus, people in ‘management’ and ‘technical’ are earning significantly higher than ‘workers’, which is considered to be the base level in the model. people in ‘services’ and ‘office’ are earning significantly lesser than the base level and although the coefficient corresponding to ‘sales’ is negative, it is not significant, i.e. they are earning the same as ‘workers’ on average.

From this, we can think of combining these levels according to wage. People in ‘management’ and ‘technical’ are in one group, can be considered as high income job group, similarly ‘services’ and ‘office’ in low income job category and ‘sales’ and ‘workers’ in medium income job category. We call this new variable ‘job’

The boxplot clearly shows the hierarchy.



Considering again a linear model of ‘wage on ‘job’ gives

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.0144 0.3328 21.077 < 2e-16 \*\*\*

job2 1.2487 0.4621 2.702 0.00711 \*\*

job3 4.9900 0.4860 10.268 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

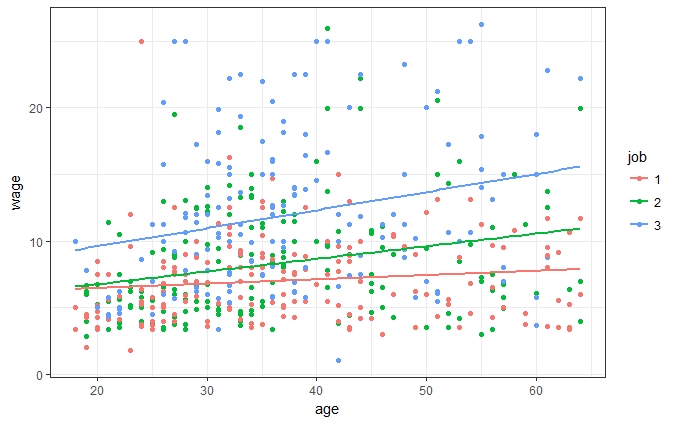
Residual standard error: 4.465 on 530 degrees of freedom

Multiple R-squared: 0.1755, Adjusted R-squared: 0.1724

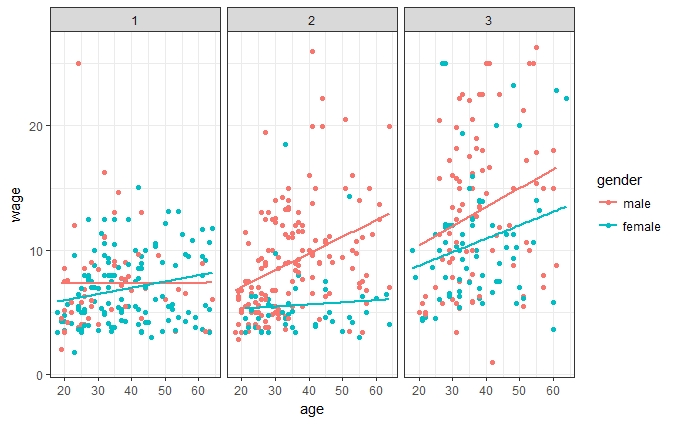
F-statistic: 56.41 on 2 and 530 DF, p-value: < 2.2e-16

This validates the plot.

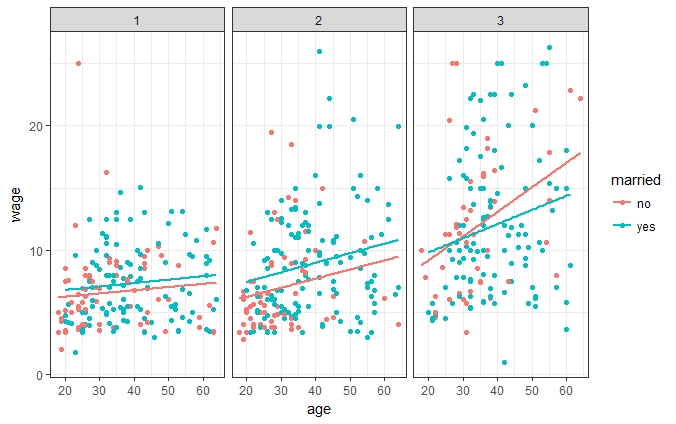
Again, considering them with ‘age’ shows that job category 3 pays much more wage at all ages, and job category 2 although pays the same as job category 1 initially, but their growth in wage with age is much faster than job category 1.



Incorporating this info into the previous plots gives

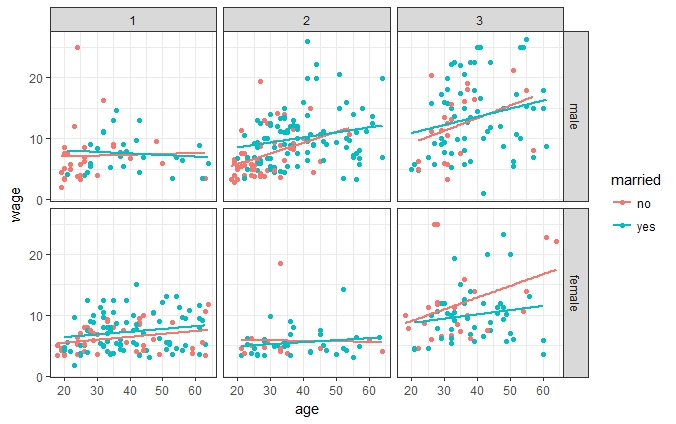


This clearly shows that gender gap is mainly due to job category 2 and 3 and growth of income in job category 2 is almost stagnant.

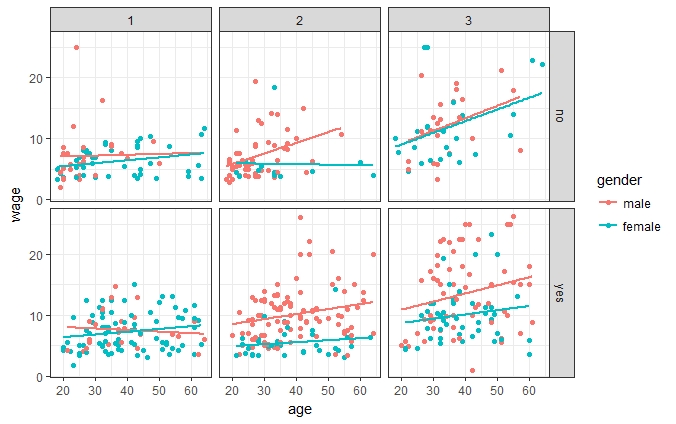


In job category 1 and 2, married people are earning more than unmarried ones irrespective of age. But in job category 3, after 30 years of age, this trend reverses.

Again dividing them further into finer groups, we observe wage gap in males due to marital status only at job category 2, and for females only at job category 3.



Here, wage gap is very prominent in job category 2 for both married and unmarried people and also in job category 3 for married people only.



Thus, overall we can conclude that,

1. Wage discrepancy is mainly observed in job category 2 and 3.
2. In job category 2, for all age groups, males are earning more than females and married males are earning more than unmarried ones.
3. In job category 3, there is not much wage gap between males and unmarried females, but married females are earning significantly lower than them.