Assignment 3

Data Visualization STAE04

Yiying (Linda) Ren

2022-10-13

1 Data preparation

In this assignment, we will use the data set called Wages in the *Ecdat* package.

This data set was collected for panel study of income dynamics and actually represents matched data: panel data from the years 1976 to 1982 (a duration of 7 years of study) and is organized such that every 7 consecutive observations belong to a separate individual. Before we present the summary of the data set, we will have to do a bit of data wrangling first. Firstly we can encode the information properly with adding two new variables *id* and *year* and secondly we add another new variable *wage* (in its original form before the logarithm transformation into *lwage*) to the data set.

Data source: Cornwell, C. and P. Rupert (1988) "Efficient estimation with panel data: an empirical comparison of instrumental variables estimators", Journal of Applied Econometrics, 3, 149–155.

2 Task 1

2.1 Summary of the data set and its variables

A short summary of data set: this data set is panel data of individual wages of 595 individuals in the United States from 1976 to 1982. In total there are 4165 observations (595 attendees X 7 years) and 15 variables, among which there are several categorical variables classified as factors. In this assignment, we will look at data on wages to try to ascertain which of these factors may be used to explain variation in wages among workers.

Table 1: Summary of variables of data set wages after data wrangling (collected in USA from 1976 to 1982).

Variable	Data Type	Class	Description
exp	numerical	integer	years of full-time work experience
wks	numerical	integer	weeks worked within the year
bluecol	nominal	factor	blue collar?
ind	nominal	factor	works in a manufacturing industry?

Variable	Data Type	Class	Description
south	nominal	factor	resides in the south?
smsa	nominal	factor	resides in a standard metropolitan statistical are?
married	nominal	factor	married?
sex	nominal	factor	a factor with levels (male,female)
union	nominal	factor	individual's wage set by a union contract?
ed	numerical	integer	years of education
black	nominal	factor	is the individual black?
lwage	numerical	continuous	logarithm of wage
id	numerical	integer	id of the attendees
year	numerical	integer	the year data corresponded to
wage	numerical	continuous	monthly wage in US dollars

2.2 Data exploration

Let's explore this data with visualization. Firstly we make a simple plot about the development of the wages among the people in this data set across the years and study its pattern. Here we use *lwage* (the logarithm form of wages) for a better visualization of the data, we group the data by *id* and present the development of wage for each individual through the years using lines to clear visualization. We also use ´opacity´ to overcome the overlap in the plot.

In plot 1, we can see there is a clear upward trend in the data which indicates a steady increase in the wage on the y-axis through the years from 1976 to 1982; also we can see that there is greater spread (heteroscedasticity) in the data as years increase, and this might be caused by the great recession started in 1979. Also for the wage in 1976 - 1977, it seems that nobody in this data set earned more than $1000. This seems very strange and the reason to that particular phenomenal is unknown to us. It might be due to some data collection error or formatting error while inputting the data, however, we are not certain.

3 Task 2

Here we will investigate whether or not longer education leads to higher wages in the data, based on only the observations from year 1982.

As we can see in plot 2, that there is a clear upward trend in the wage as the increasing of the length of education. Also we can tell that only a few people had less than 6 years of education and a great deal of people had around 12 years of education which means high school level. People who had longer length of education sometimes could have lower wages than others who had shorter length of education. This might be due to other factors such as working experience and worked weeks in that year etc.

To get a better idea of the difference in the median and the spread (IQR) of the monthly wages for different lengths of education (categories), we make a box plot to better illustrate the differences between the categories for wages in 1982.

It seems that box plot gives us an even clearer idea about the data's (log scale of wage) median and IQR and reals the outliers in each category for the lengths of education.

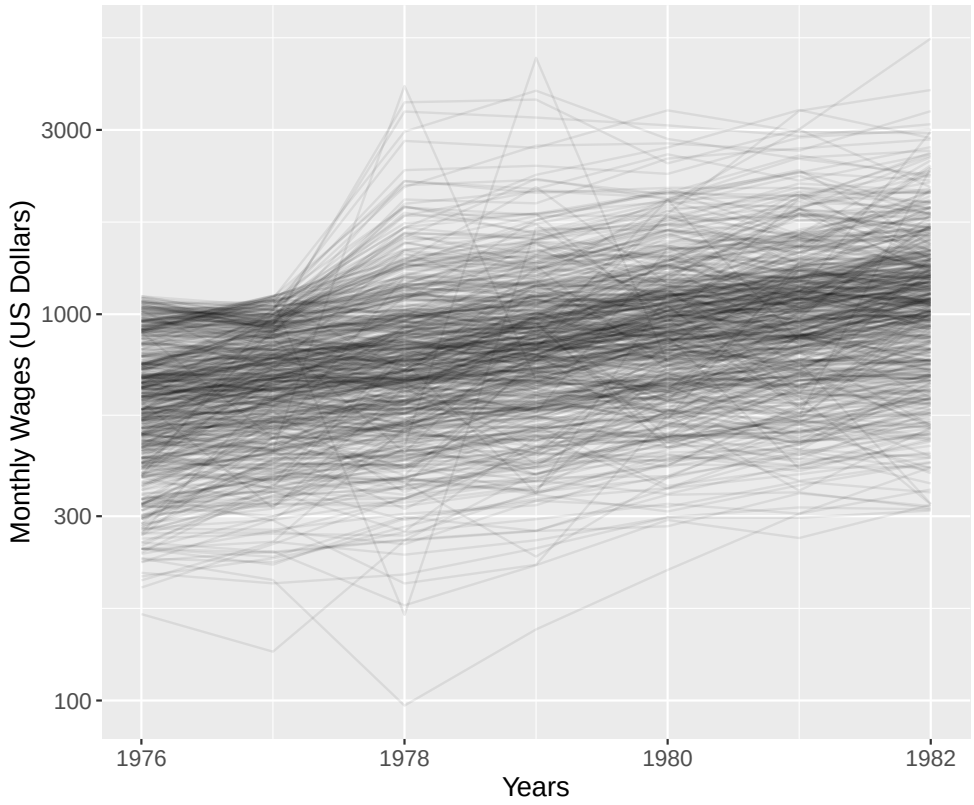

Figure 1: Monthly wages (in its logarithm form) development through the seven years in the US (Data set Wages from 1976-1982)

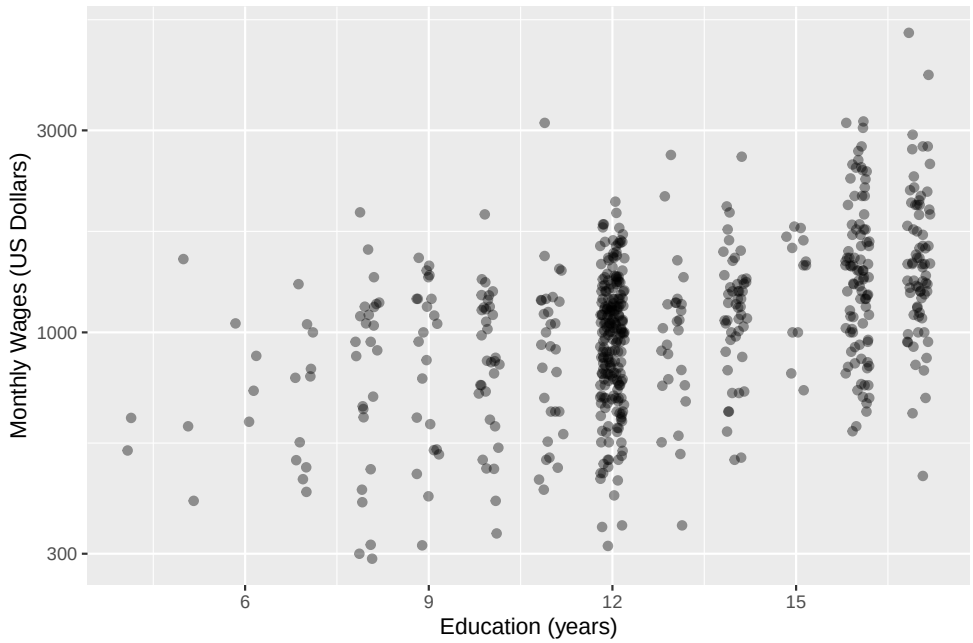

Figure 2: Scatter Plot: The length of education vs monthly wages (in its logarithm form) in the US for year 1982 (data set Wages)

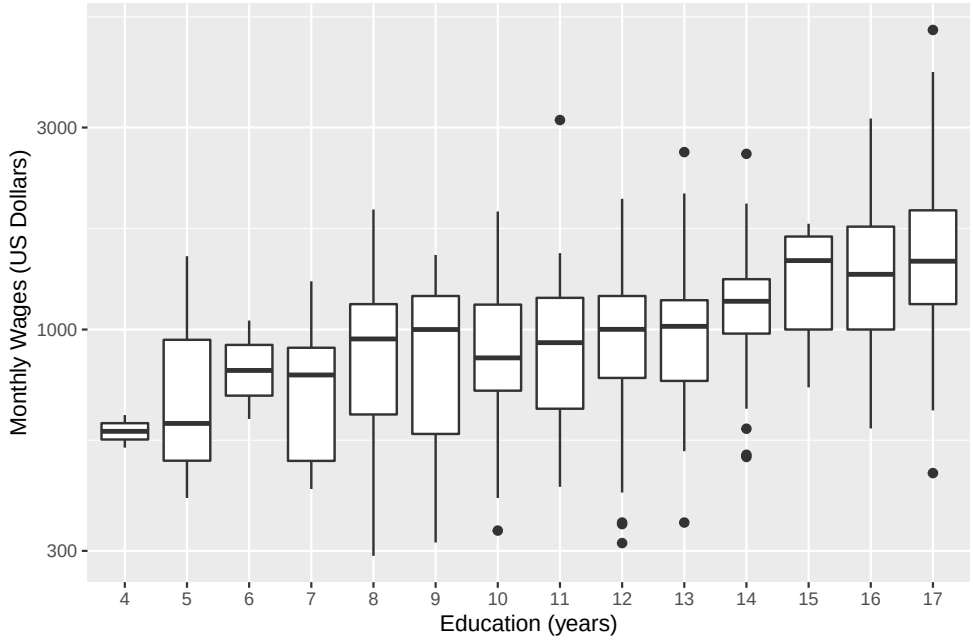

Figure 3: Box Plot: The length of education vs monthly wages (in its logarithm form) in the US for year 1982 (data set Wages)

The box plot shows that the median increases irregularly with the increase of years of education, and there is quite a big jump after completing 15 years of school which suggests that a college degree boosts one's wage. There is however higher variability here, as indicated by the larger boxes, whisker lengths and outliers. There is also a decrease in wages at year 16 and 17 in comparison to 15, this is likely that some continued their studies and worked part-time or only landed low-paying jobs as interns or research assistants, while others came up with great ideas that they could monetize, or were offered lucrative jobs.

4 Task 3

Now, let's take a look at the relationship between wage and some of the other variables in the data set. Here we will convert *wage* from a continuous variable into a categorical variable (factor) with three levels: low, medium and high. This is mostly to accommodate the purpose of practicing the visualization of categorical variables.

We could compare the wage levels between woman and men and also between people with and without a dark skin color in the US from 1976 to 1982 with this following plot.

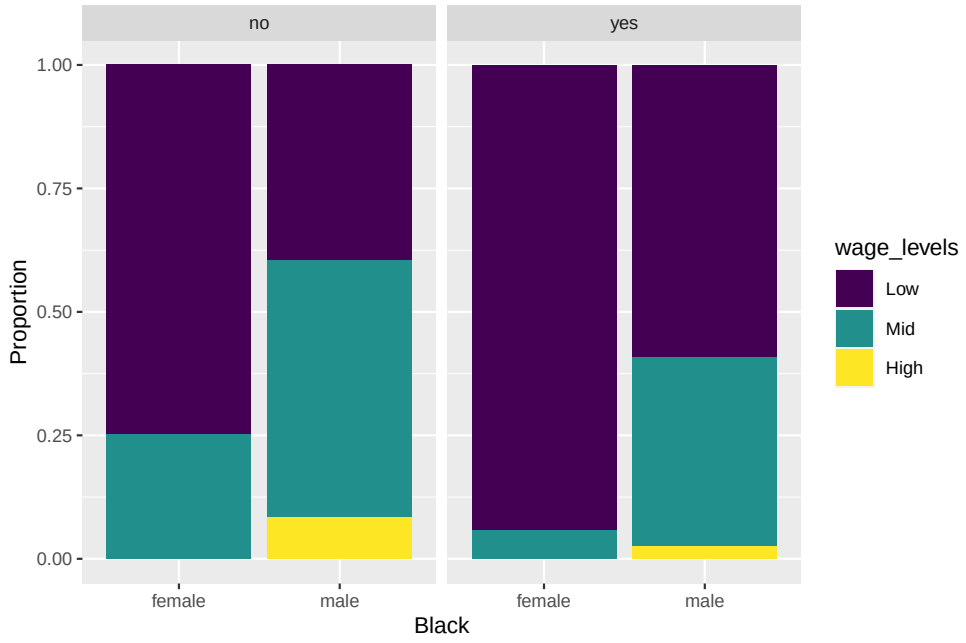

Figure 4: Wage Level comparison: grouped by gender and skin color (US from 1976 to 1982)

As plot 3 shows that in the US from 1976 to 1982, people without dark color had higher wages than people with dark color and this is true for both men and women; furthermore that men in general had higher wages than woman in spite of the skin color, which means that black men have higher wages than both black and non-black women. This plot shows rather clearly that according to data set *Wages*, back in

1976 - 1982 the prejudice and inequality in gender and race had a great impact on the income level in the US.

The reason I choose this stacked bar plot is I think it is interesting to compare the different income levels between men and women and also between different races. And I think this stacked bar plot serves this purpose rather well. It is simple and gives a clear intuitive idea about the differences between the categories.

Extra: to see if years of experience would affect on the wage.

Figure 5: Wage Level comparison: grouped by experiences (US from 1976 to 1982)