Assignment 3

STAE04 - Statistics: Data Visualisation

Thi Kim Hong Nguyen

2024-02-18

This Assignment will work with two variables at a time as well as categorical data. In this Assignment, we will examine wage data to identify potential factors contributing to the variability in earnings among individuals. The dataset utilized for this analysis is the Wages dataset sourced from the Ecdat package.

Then we will run the code to make the data set available in R and format the data set. This data set actually represents matched data: panel data from the years 1976 to 1982 and is organized such that every 7 consecutive observations belong to a separate individual. Here, we simply repeat each value of the vector 1:595 seven times, and store this new variable as id and add another variable called year, which is simply the vector 1976:1982 repeated 595 times.

```r
install.packages("Ecdat")
```

```
## package ’Ecdat’ successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\DellVostro5590\AppData\Local\Temp\Rtmp86mBRq\downloaded_packages
```

```r
library(Ecdat)
data(Wages)
head(Wages)  #Let's take a look at the data set
```

```
##   exp wks bluecol ind south smsa married  sex union ed black
## 1   3  32      no   0   yes   no     yes male    no  9    no
## 2   4  43      no   0   yes   no     yes male    no  9    no
## 3   5  40      no   0   yes   no     yes male    no  9    no
## 4   6  39      no   0   yes   no     yes male    no  9    no
## 5   7  42      no   1   yes   no     yes male    no  9    no
## 6   8  35      no   1   yes   no     yes male    no  9    no
##   lwage
## 1  5.56
## 2  5.72
## 3  6.00
## 4  6.00
## 5  6.06
## 6  6.17
```

```r
dim(Wages)  #Check data set's dimensions
```

```
## [1] 4165   12
```

```r
#formatting data set
wages_formatted <-
  Wages %>%
  mutate(
    id = as.factor(rep(1:595, each = 7)),
    year = rep(1976:1982, times = 595)
  )
```

1 Task 1: Getting the Data

I used the Wages dataset is part of the Ecdat package in R. The Wages dataset from the Ecdat package provides information on wages and various factors related to individual workers. It includes variables such as income, education, experience, and other relevant factors that can be used for the analysis of wage disparities and patterns among workers. The dataset is structured to facilitate research and exploration of the relationships between different variables and wages in the context of labor economics.

From 1976 to 1982 a panel of 595 observations of individual workers were collected, annd organized such that every 7 consecutive observations belong to a separate individual.Therefore, there are 4165 observations in total with 12 Variables.

```r
wages_formatted <-
mutate(wages_formatted, wage = exp(lwage)) #add wage variable as their original scale
```

I also added a new variable to the data set containing the wages at their original (non-transformed) scale, called wage. The description of variables data set are the following below:

Table 1: The variables of the data set

Attribute	Description
exp	years of full-time work experience
wks	weeks worked
bluecol	blue collar?
ind	works in a manufacturing industry ?
south	resides in the south ?
smsa	resides in a standard metropolitan statistical are ?
married	married ?
sex	a factor with levels (male,female)
union	individual's wage set by a union contract ?
ed	years of education
black	is the individual black ?
lwage	logarithm of wage
id	id

Attribute	Description
year	year
wage	wage per month

First, let's examine how the wages of individuals in this data set have evolved.

```
# Create a line plot connecting individuals over the years
ggplot(wages_formatted, aes(x = year, y = wage)) +
  geom_line(aes(group=id)) +
  labs(x = "Year",
       y = "Wage per Month")
```

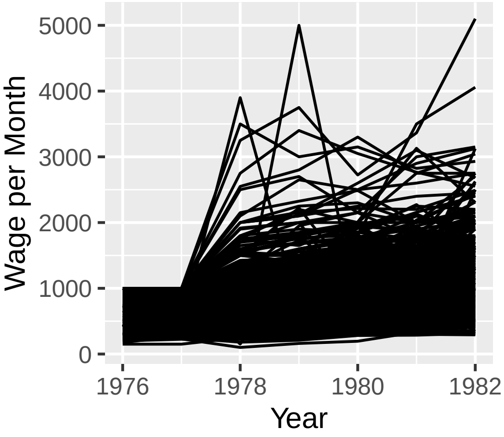

Figure 1: Individual Wages Over Years

Overall, wages show an increasing trend from 1976 to 1982, reflecting economic development. Nevertheless, certain individuals exhibit a notable decline in the period between 1978 and 1980, attributable to the global economic crisis.

2 Task 2: Exploring the Depths of the Data

In this task, I will emulate the approach taken in the initial report and explore the hypothesis suggesting a positive correlation between longer education and higher wages. To facilitate data visualization, I will narrow down the data set to exclusively encompass observations from the year 1982.

```
# Create a line plot connecting individuals over the years
library(dplyr)
wages_1982 <- wages_formatted %>% #filter observations only in 1982
  filter(year == 1982)
# Create a plot explores the relationship between years of education and wages
ggplot(wages_1982, aes(as.ordered(ed),wage)) +
  geom_boxplot() +
  labs(x = "Years of Education", y = "Wages per month")
```

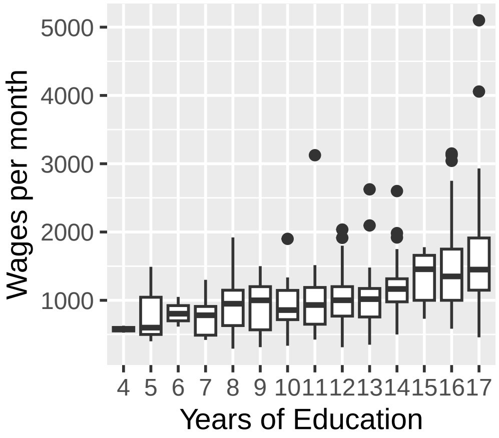

Figure 2: Individual Wages Over Years

Now, We can actually see a slight improvement in wages per month over the years of education, at least in terms of the median values. The plot illustrates that an additional year of schooling produces a 10% wage gain. Notwithstanding, it should be noted that there is no significant difference in income between two years of consecutive experience.Furthermore, it does not account for the possible correlation of explanation variables with the individuals effects. The analysis can be skeptical due to it might not account for all relevant variable influencing wages such as job type, industry or geographical location. For example, we can consider into the outlines from above 10 years experience. There are some excellent individuals that have wages significant different from the average.

3 Task 3:

In this task, we will examine the relationship between wage and other variables in the data set using wage_level with three categories (“Low”, “Mid”, “High”) based on the wage values. First, take a look at the relationship between wage level and sex.

```
# Transform the wage variable into a categorical factor
wages_with_levels <-
  wages_formatted %>%
  mutate(
    wage_levels = cut(
      wage,
      c(0, 750, 1500, max(wage)),
      labels = c("Low", "Mid", "High"),
      ordered_result = TRUE
    )
  )
#create bar plot between wage and gender
ggplot(wages_with_levels, aes(sex, fill = wage_levels)) +
  geom_bar(position = "fill") +
  labs(y ="Proportion")
```

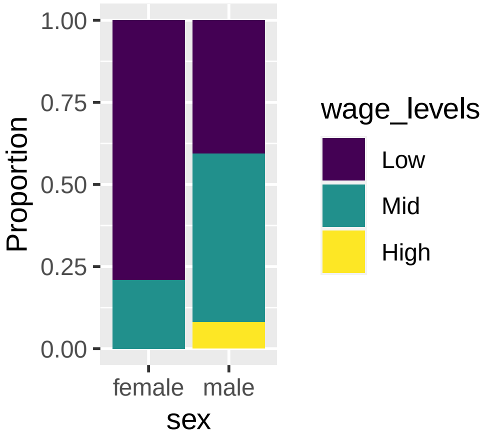

Figure 3: Distribution Wages Between Male and Female 1976 -1982

It is evident that men generally have higher salaries than women. A significant portion of men falls into either the low or high-wage categories, with approximately 20% having high wages. In contrast, there are no women with high salaries, and the majority of them earn lower wages. Parenthood is a contributing factor to the widening gender pay gap. Women between the ages of 25 and 44 who are mothers are less likely to participate in the labor force compared to their childless counterparts in the same age group. Furthermore, when employed, these mothers tend to work fewer hours per week. In contrast, fathers are more inclined to engage in the labor force and dedicate more hours per week to work compared to men without children at home.