Assignment 3

STAE04/DABN19: Data Visualization

Dirk Baars

2021-10-14

1 Task 1

The data set contains panel data of 595 observations, that was collected between 1976 and 1982 in the United States. The total numbers of observations are: 4165. The Wages data set contains 12 variables, whereby one variable is added, by taking the inverse log of lwages.

The description of the variables are as follows:

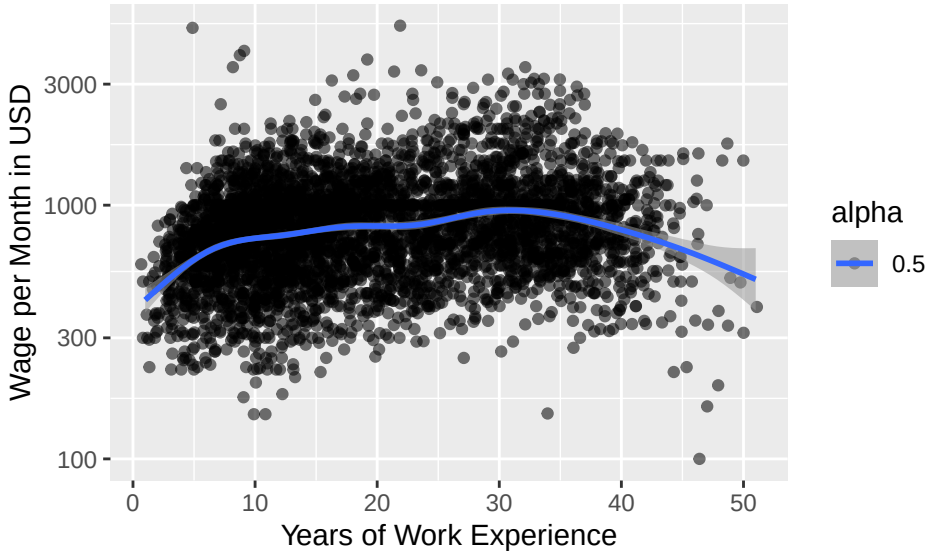
Name	Transformation	Measurement	Description
exp	—	years	years of full-time experience
wks	—	weeks	weeks worked
bluecol	—	yes, no	blue collar?
ind	—	yes, no	works in a manufacturing industry ?
south	—	yes, no	resides in the south ?
smsa	—	yes, no	resides in a standard metropolitan statistical are ?
married	—	yes, no	married ?
sex	—	male, female	sex of the person
union	—	yes, no	individual's wage set by a union contract ?
ed	—	years	years of education
POC	—	yes, no	is the individual person of color ?
lwage	log transformation	logarithm	logarithm of wage
wage	—	—	wage in month

The graph below shows a relation between the years of experience, and the wage per month. A geom_point() is used to visualize the graph.To prevent overlapping, the position is jittered, and opacity (alpha = 0.5) is adjusted. Also, a logarithmic scale is adjusted to give a better overview of the data. Also, a line is added by using geom_smooth(). The graph shows that until approximately 30 years of working experience, the wage seems to increase, and after that the wages slowly decreases. The graph has no clear pattern, whereas the variability seems high.

```
Wages1 %>%
ggplot(aes(exp, wage, alpha = 0.5)) +
geom_point(position = position_jitter()) +
```

```
  labs(
    x = "Years of Work Experience",
    y = "Wage per Month in USD"
  ) +
  geom_smooth() +
  scale_y_log10()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Task 2

The graph below shows a relation between the logarithmic wage, and the weeks worked. The weeks worked is filtered, so only >=40 weeks is taken into account. A geom_point() is used to visualize the graph.To prevent overlapping, the position is jittered, and opacity (alpha = 0.5) is adjusted. Also, a the logarithmic wage is used to give a better overview of the data. Also, a line is added by using geom_smooth().

In the figure it shows no clear pattern, whereby wage remains fairly stable in comparison to the weeks worked.

```
Wages1 %>%
  filter(wks >= 40) %>%
  ggplot(aes(wks, lwage, alpha = 0.5)) +
  geom_point(position = position_jitter(width = 1, height = 1)) +
  geom_smooth() +
  labs(
    x = "Weeks Worked",
    y = "Log of montly Wage in USD"
  )
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

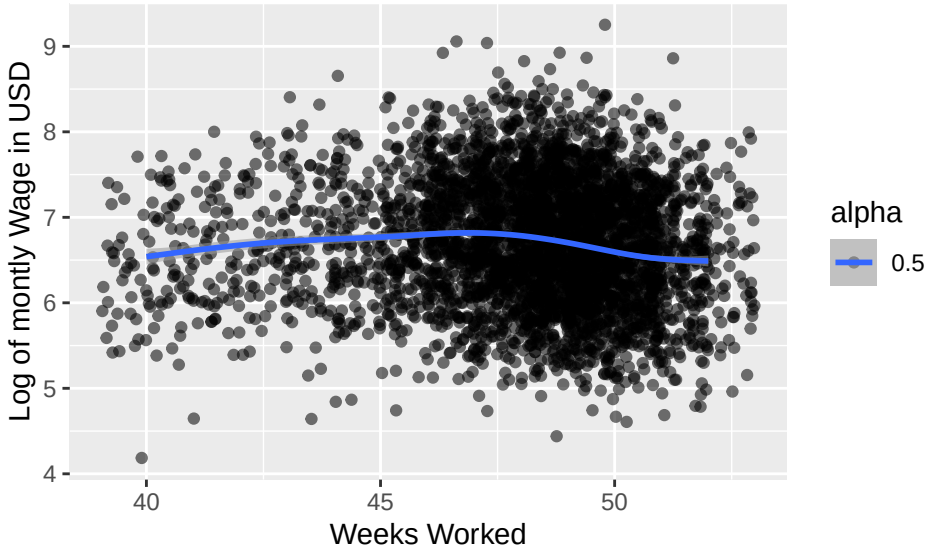


Figure 1: The amount of weeks worked has a positive relation to wage. Points are jittered

2 Task 3

When plotting sex against wage levels in a mosaic table, it is visible that there are more males in the data, but also that there are no females earning high wage and that females proportionally earn more often a low wage level.

```
Wages2 %>%
  ggplot() +
geom_mosaic(aes(x=product(sex), fill = wage_levels)) +
  labs(
    y = "Wage levels",
    x = "Sex of the individual"
  ) +
  guides(fill = guide_legend(title="Wage levels"))
```

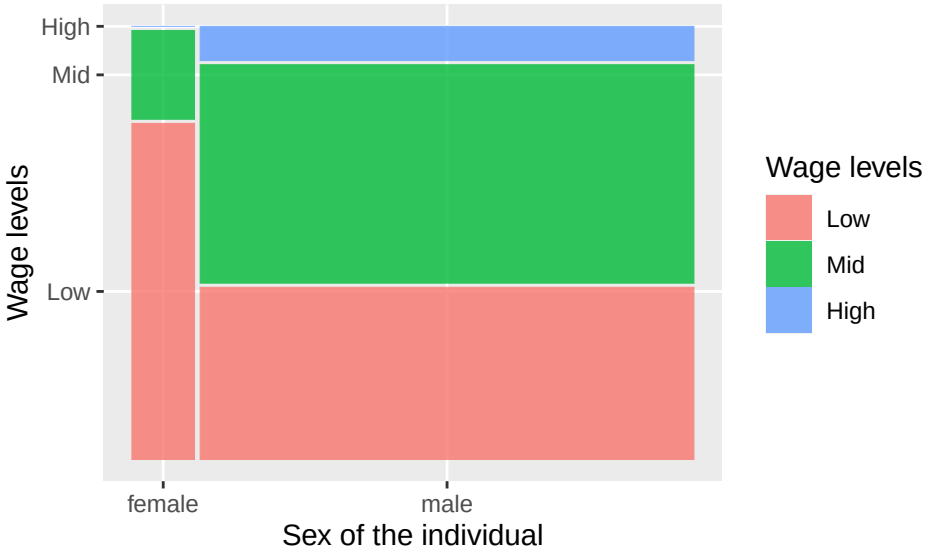


Figure 2: Wage levels of the two sexes