Initial Data Analysis

Shivangi Sinha

30th January 2019

Abstract

The uswages data frame has 2000 rows and 10 columns. Weekly Wages for US male workers sampled from the Current Population Survey in 1988.

Variables

The variable are as follows

- Real weekly wages in dollars (deflated by personal consumption expenditures - 1992 base year)
- Years of education
- Years of experience
- Race
- 1 if living in Standard Metropolitan Statistical Area, 0 if not
- 1 if living in the North East
- 1 if living in the Midwest
- 1 if living in the West
- 1 if living in the South
- 1 if working part time, 0 if not

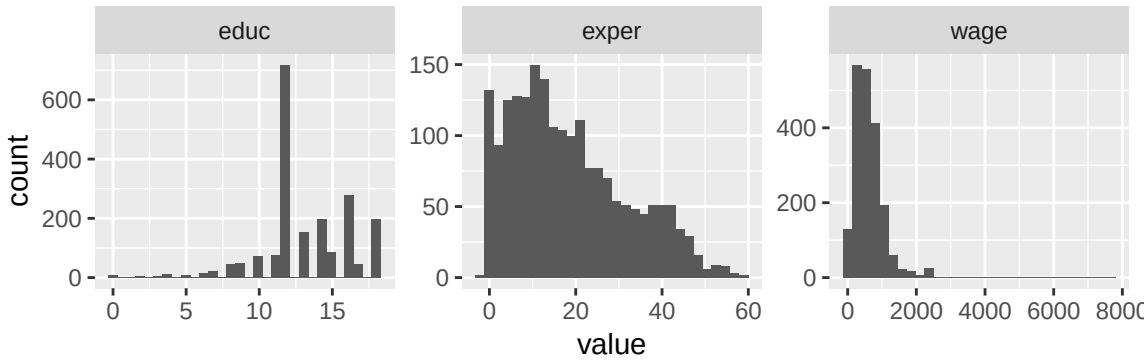
Descriptive Statistics

```
##       wage              educ           exper      
##  Min.   :  50.39   Min.   : 0.00   Min.   :-2.00  
##  1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00  
##  Median : 522.32   Median :12.00   Median :15.00  
##  Mean   : 608.12   Mean   :13.11   Mean   :18.41  
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00  
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00
```

We can see that the exper has min as -2 which is not possible. It clearly has outliers. Wage also has the max at 7716 when the mean is around 608. We should look at the plot to get more details.

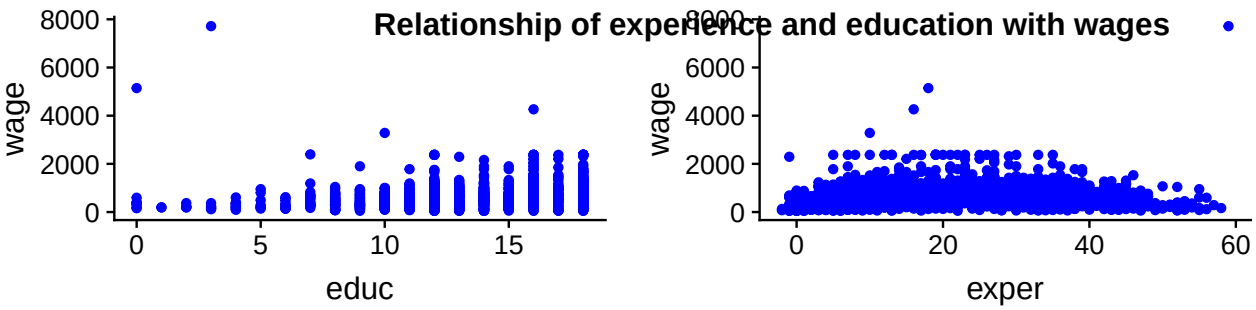
Graphical Summaries

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



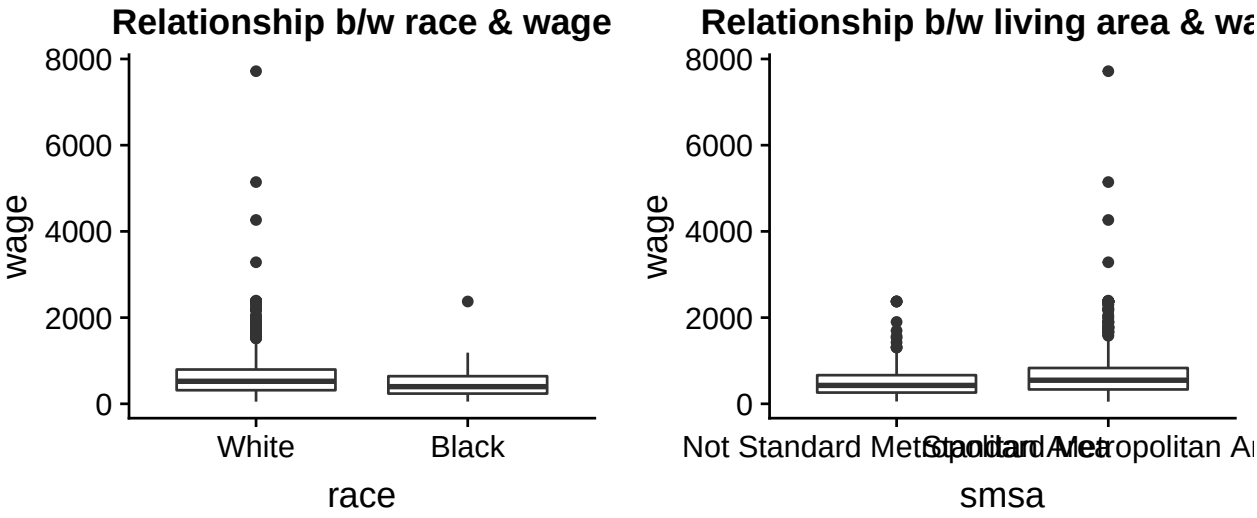
We can see that exper is skewed to the left and education is skewed to the right.

Pairwise Relationships



This is reinforces our notion that more number of years of eductions results in more wage. We can spot the outliers in this graph as well.

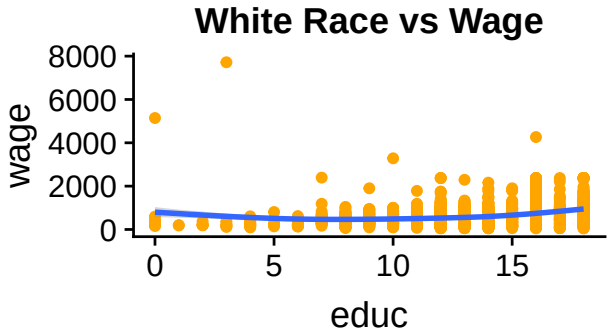
We can see how more number of experience doesn't neccesarily lead to high wages.

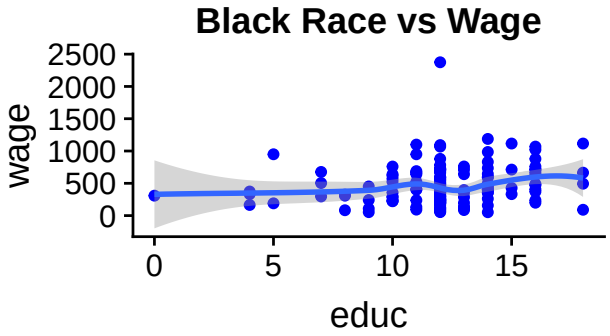


According to this the people with white race only had a slight in their wage when compared with people with black race. Again, people who lives in Statistical Metropolitan Area has only a slight increase in the wage . We can see it has more outliers present in the data.

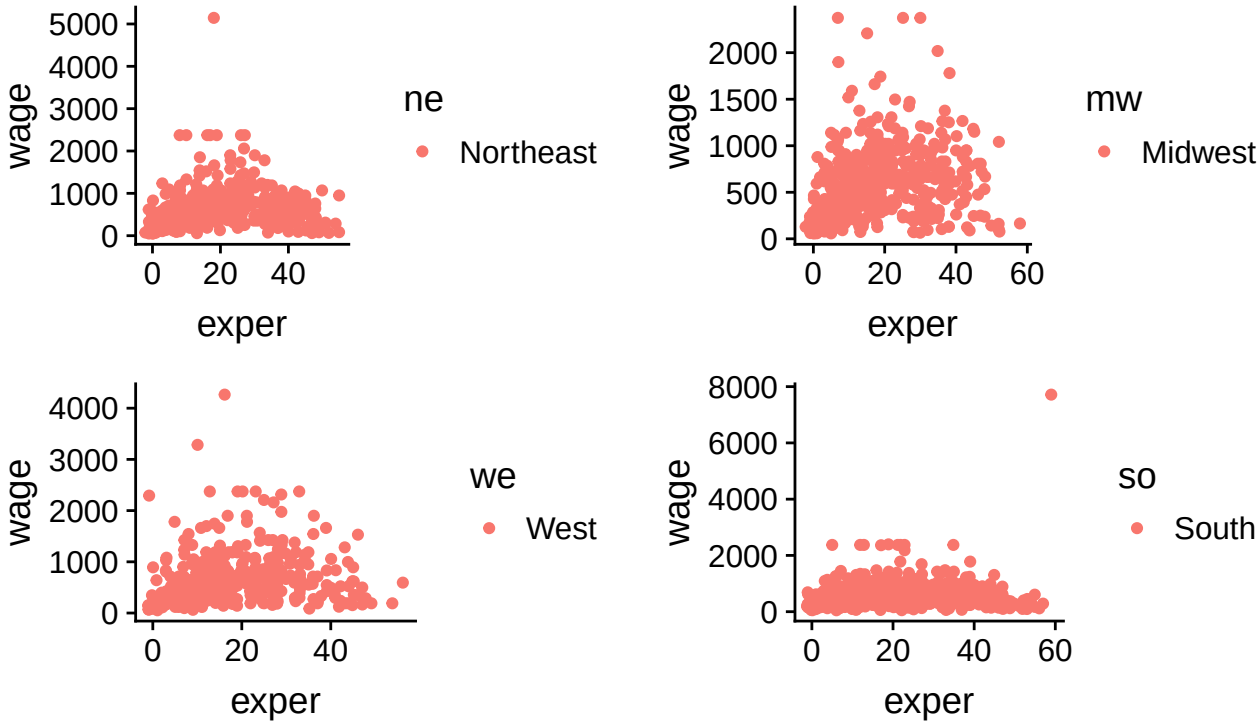
Studying Relationships between three variables.

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





We can infer from the graphs that people with white race earn significantly more than people with black race with same number of education.



It is noticeable from the graphs that people living in Northeast and South tend to get more wage with the same level of experience than their counterparts.

Takeaways

- There is a wage gap between the people of two race we are considering.
- There is sknewness and outliers present in the data that needs to be fixed.
- Relationship between wages and living area of a person needs to be more explored.