

Group 11

Wage difference between Black & White People and the reason behind it

Chen (Cici) Chen	cc4291
Sixing Hao	sh3799
Wenwen Shen	ws2561
Yuting He	yh3054
Yang Meng	ym2696

Introduction

- Data
- Histogram of Wage
- Test for normality

Data

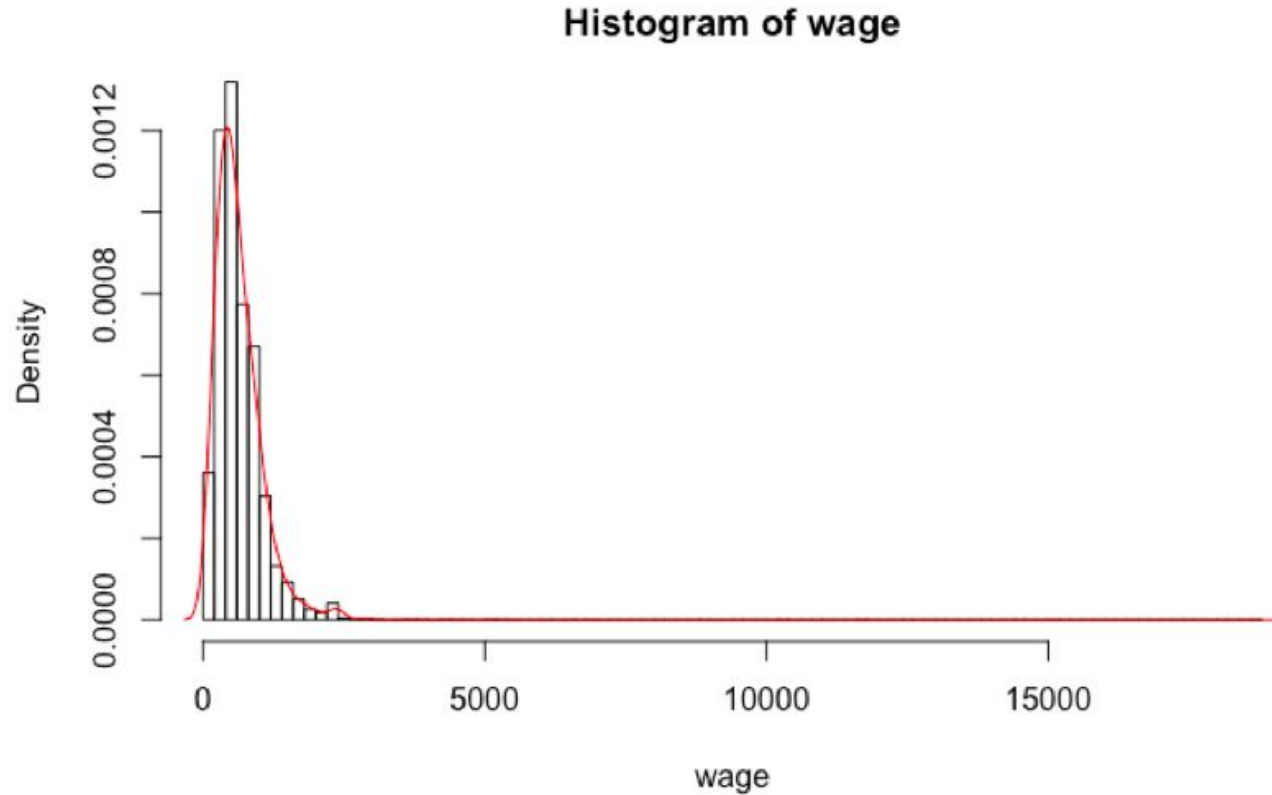
The data is roughly about 25,000 records of people between the ages of 18 and 70, and the data are taken many decades ago so the wages are lower compared to current times.

Wages ⬆	Education ⬆	Experience ⬆	Region ⬆	Race ⬆	Commute Distance ⬆	Employee ⬆
354.94	7	45	northeast	white	24.3	200
370.37	9	9	northeast	white	26.2	130
754.94	11	46	northeast	white	26.4	153
377.23	16	22	northeast	white	7.1	181
284.9	8	51	northeast	white	11.4	32
264.06	12	0	northeast	white	1	166
1643.83	14	18	northeast	white	10.2	195

↑
Response

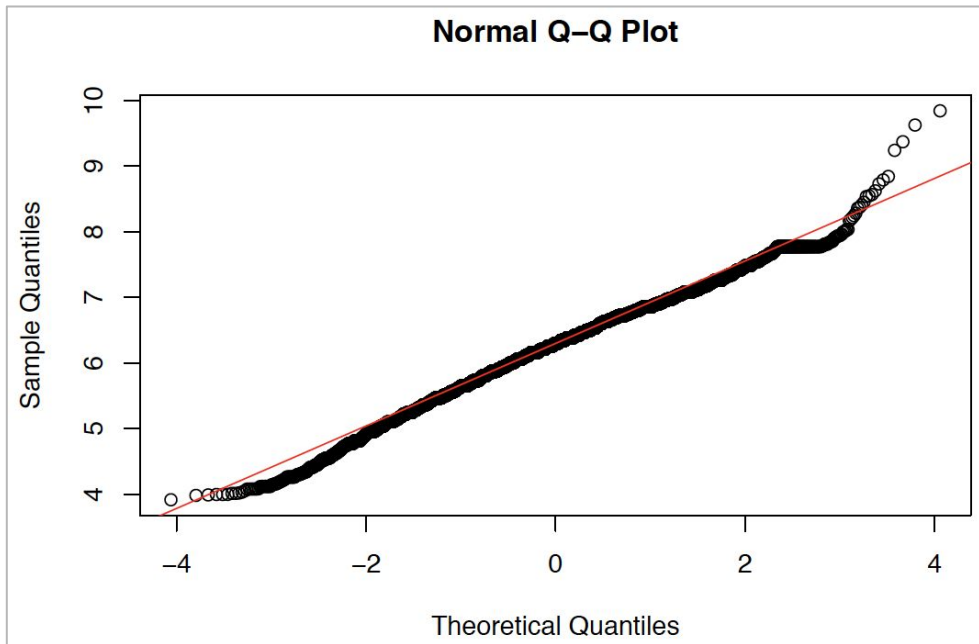
Predictors

Histogram of Wage



Test for Normality

- QQ-plot



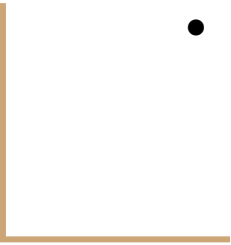
- Shapiro-wilk Test for Wage

```
## Shapiro-Wilk normality test
##
## data:  log(wage.test)
## W = 0.99418, p-value = 2.273e-13
```

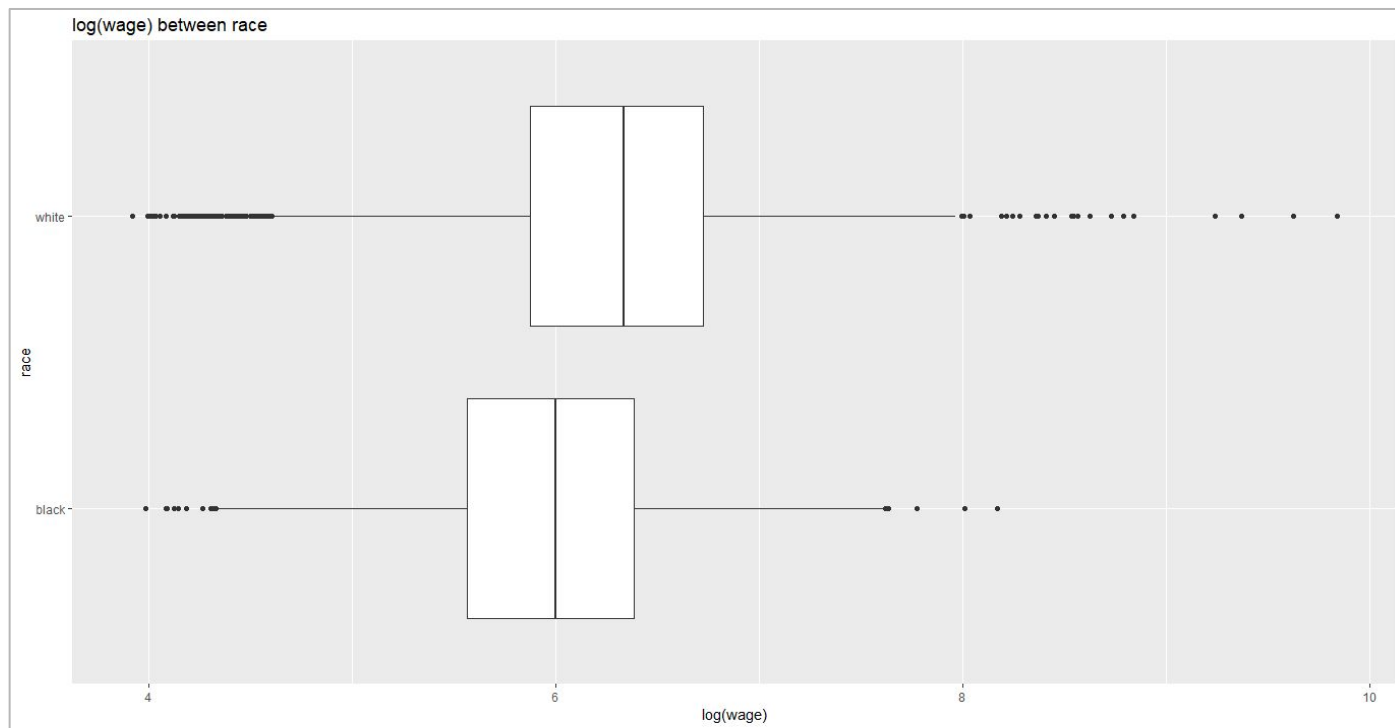
Even using **Log transformation**, it is still not normal, so that is the reason why we go to **Non-parametric** methods.



Exploratory Data Analysis

- 
- Difference of wage
 - Correlation between variables

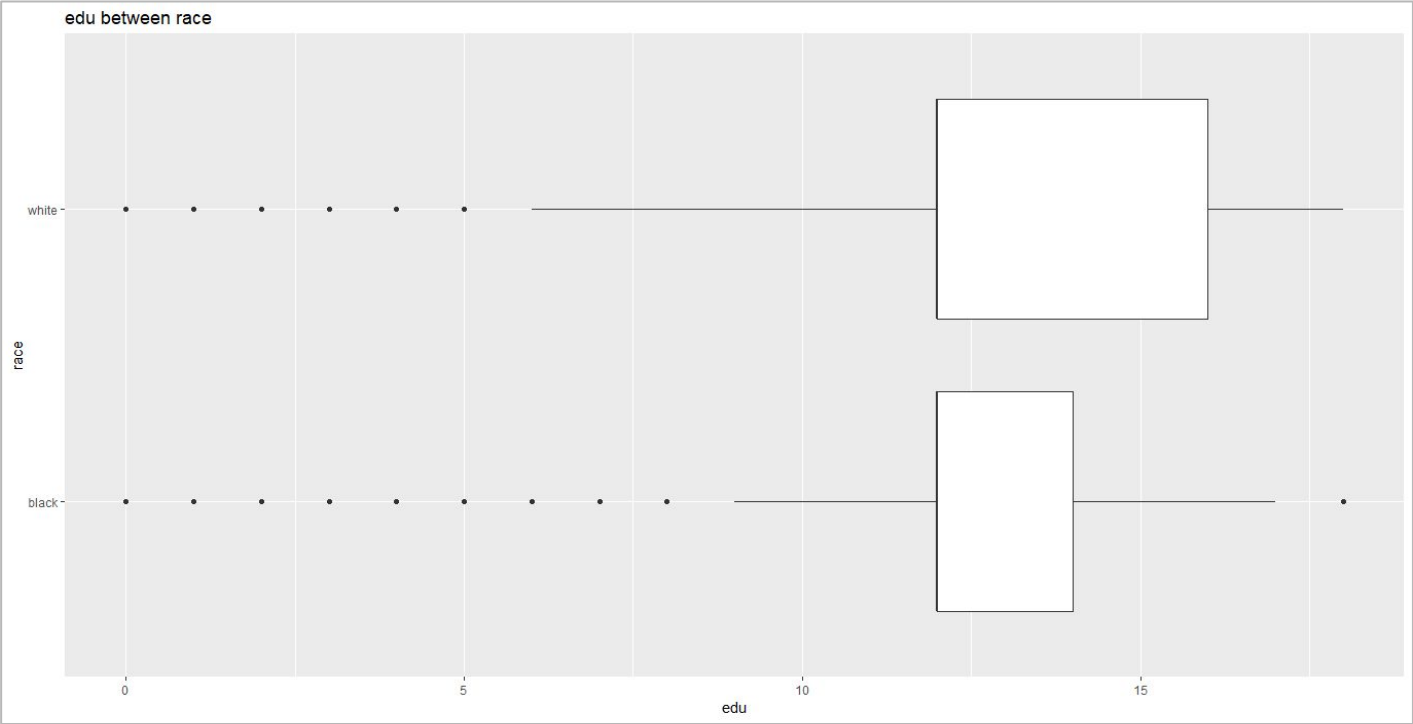
Wage



Different!



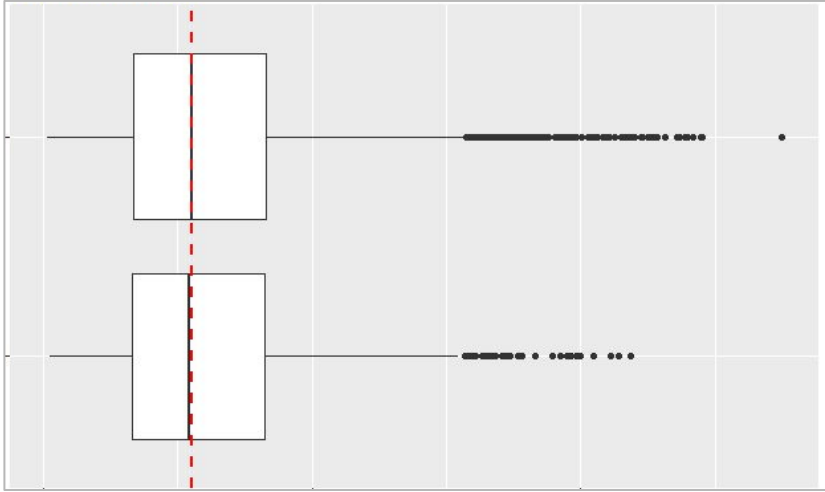
Education (in Years)



Different!

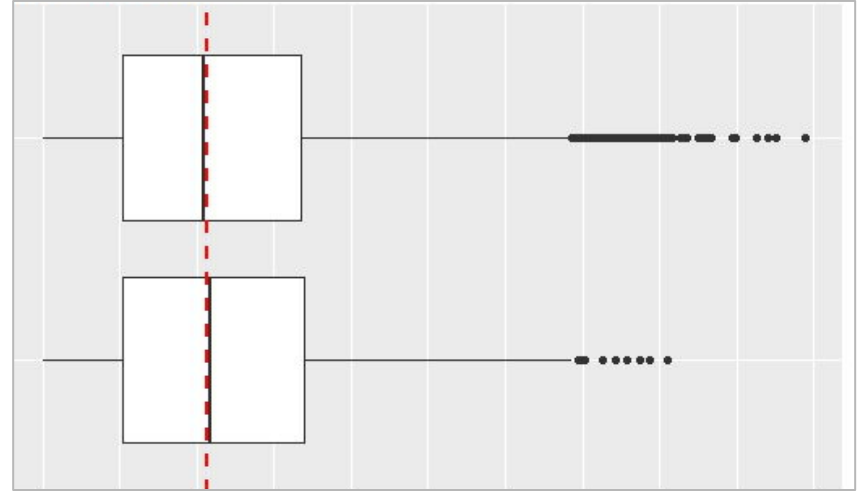


Number of Employees



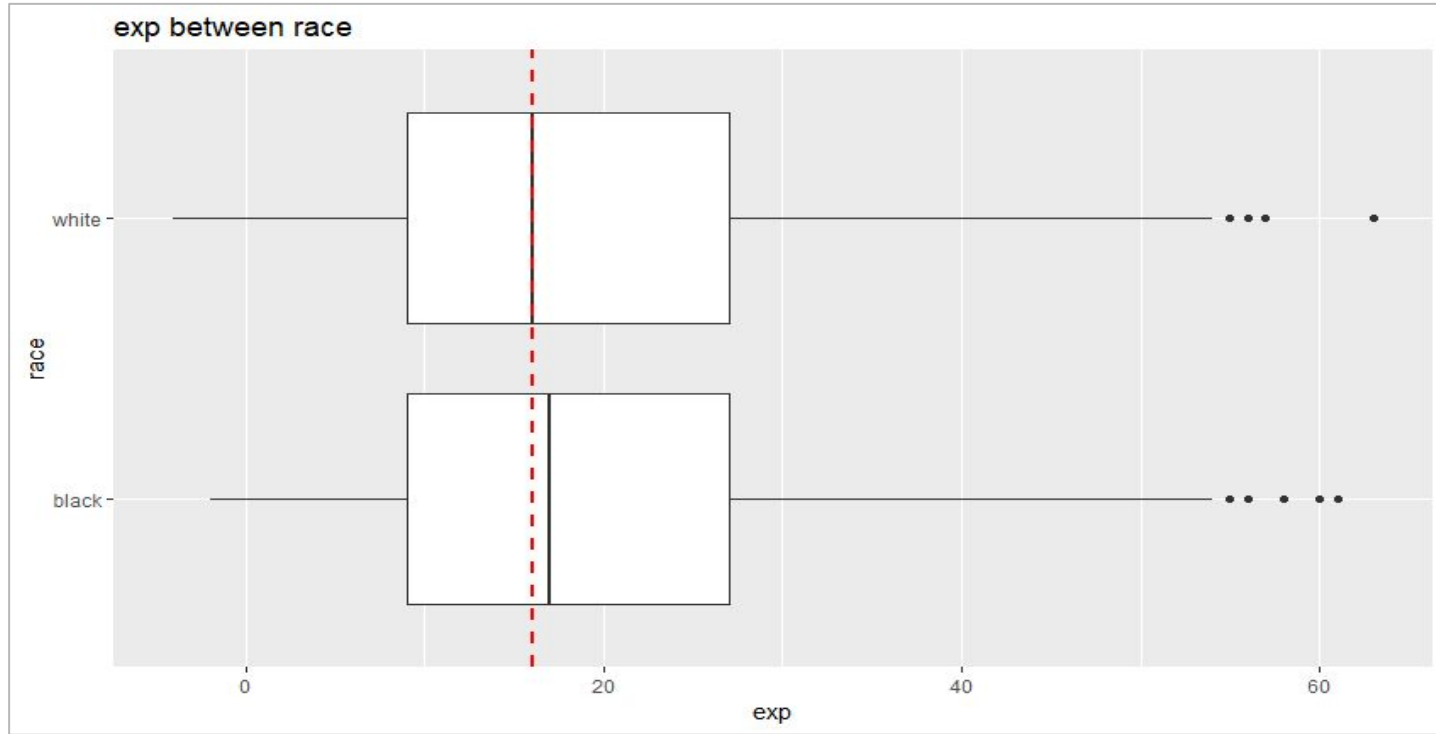
Not a big difference!

Commuting Distance



Not a big difference!

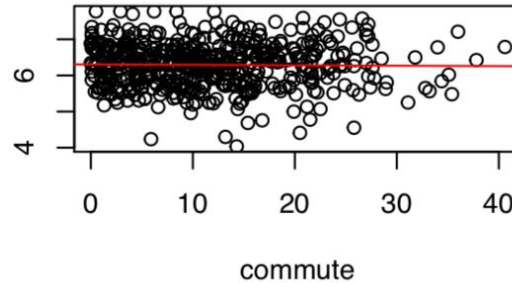
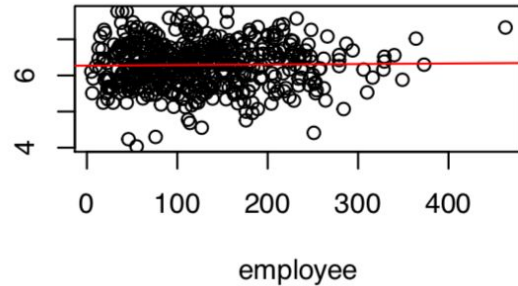
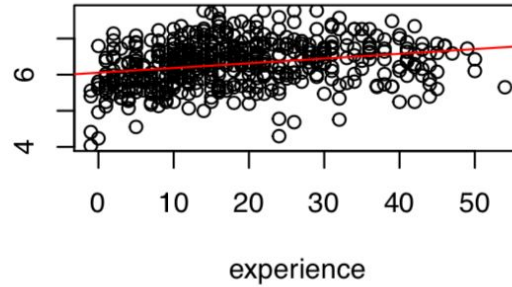
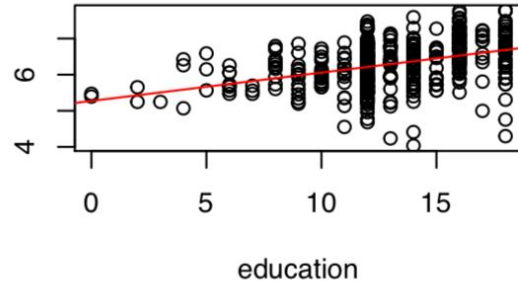
Working Experience in Years (Need Further Research)



Different!



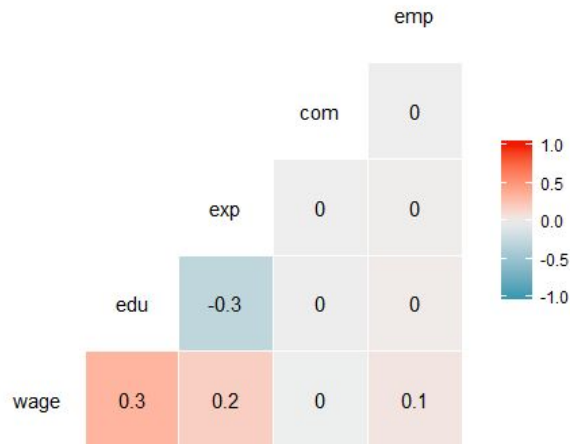
Scatter Plot



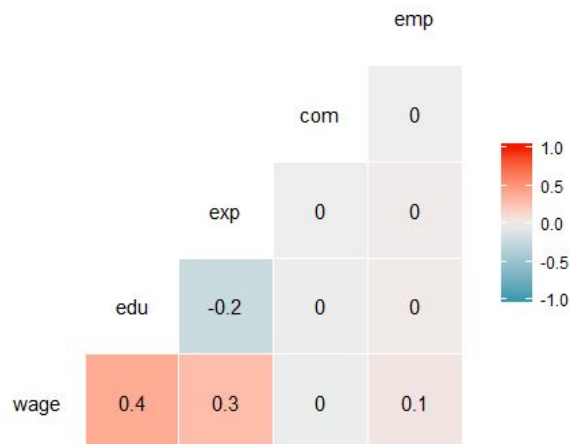
- Wage (log) against other numeric variables
- Scatter plot and best fitting line displayed similar results

Correlation

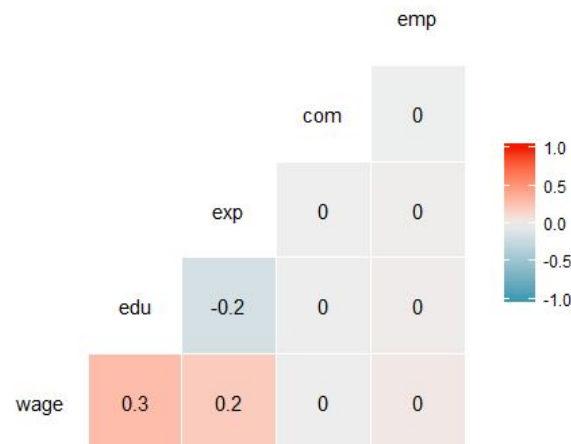
Pearson correlation r



Spearman correlation r_s



Kendall correlation r_τ



$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$$t_{corr} = \sqrt{\frac{n-2}{1-r^2}} r \sim t_{(n-2)} \quad \text{under } H_0$$

$$r_s = \frac{\sum_{i=1}^n (R_i^x - \frac{n+1}{2})(R_i^y - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^n (R_i^x - \frac{n+1}{2})^2 \sum_{i=1}^n (R_i^y - \frac{n+1}{2})^2}}$$

$$Z = \frac{r_s}{\sqrt{\text{Var}(r_s)}} = r_s \sqrt{n-1} \sim N(0, 1) \quad \text{under } H_0$$

$$r_\tau = 2 \frac{\sum_{i=1}^{n-1} V_i}{C_n^2} - 1$$

$$Z = \frac{r_\tau}{\sqrt{\text{Var}(r_\tau)}} = \frac{r_\tau}{\sqrt{\frac{4n+10}{9(n^2-n)}}} \sim N(0, 1) \quad \text{under } H_0$$

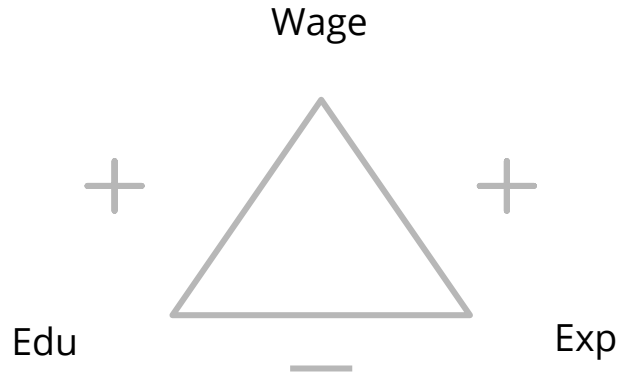
where V_i is the number of pairs (X_i, Y_i) that are concordant

Short Conclusion for EDA

Boxplot

Variables	Wage	Education Years	Number of employees	Commuting Distance	Working years
Conclusion	Different	Different	Not a big difference	Not a big difference	Need further research

Correlation



Tests

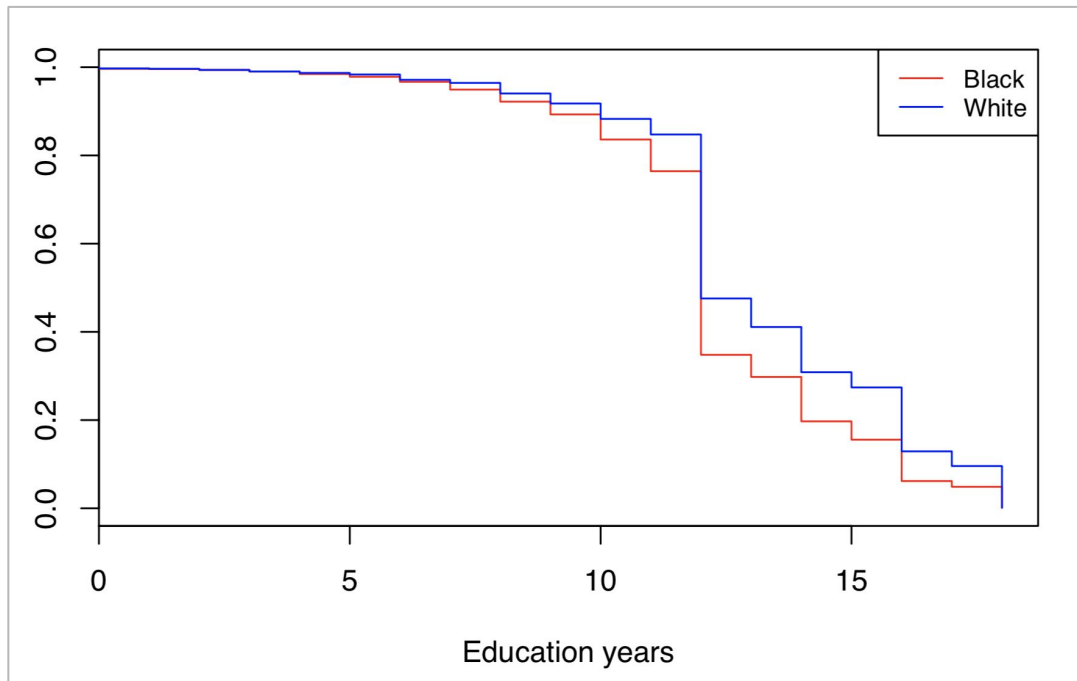
- Two-sample tests for wage
- Survival Analysis for education years
- M-H test for work experience

Two Sample Tests for Difference in Wage

	Output	Standard Version	Permutation Version
t-test	Welch Two Sample t-test data: black and white t = -22.975, df = 2952.7, p-value < 0.00000000000000022 alternative hypothesis: true difference in means is not equal	P-value ≈ 0	P-value ≈ 0
Wilcoxon Rank-sum test	Wilcoxon rank sum test with continuity correction data: black and white W = 12537000, p-value < 0.00000000000000022 alternative hypothesis: true location shift is not equal to 0	P-value ≈ 0	P-value ≈ 0
F-test	Analysis of Variance Table Response: wage Df Sum Sq Mean Sq F value Pr(>F) 1 56713844 56713844 275.6 < 0.00000000000000022 *** Residuals 20237 4164404694 205782	P-value ≈ 0	P-value ≈ 0
Kruskal-Wallis test	Kruskal-Wallis rank sum test data: wage by race Kruskal-Wallis chi-squared = 446.68, df = 1, p-value < 0.00000000000000022	P-value ≈ 0	P-value ≈ 0

Different!

Survival Analysis for Education (in Years)



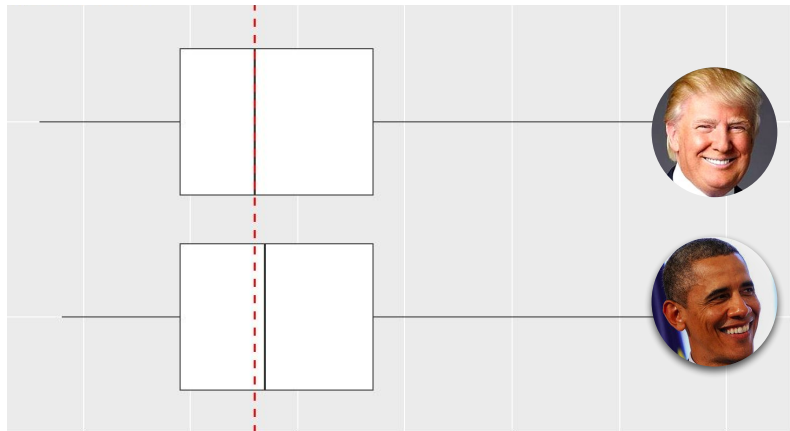
Likelihood ratio test= 143.9 on 1 df, $p < 2e-16$
Wald test = 155.9 on 1 df, $p < 2e-16$
Score (logrank) test = 157.1 on 1 df, $p < 2e-16$

Censored  Working

Uncensored  Still studying

Black People receive less education compared to **White People**.

Chi-square test of Experience between Races

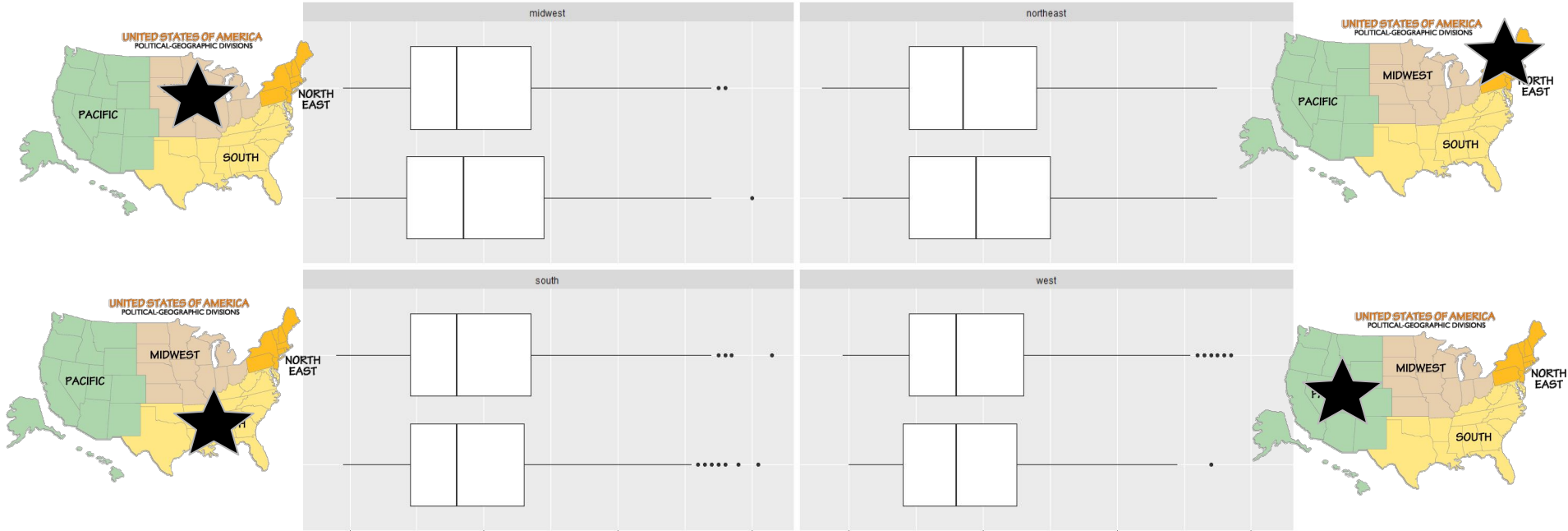


	Inexperienced	Experienced
Black	954	980
White	9333	8972

Basic Approach

- Experienced = (Working years > 16)
(16 years is the median)
- Chi-square test leads to the same result by **p-value=0.1728**
- So, the proportion of experienced workers in two races has no difference in this test.
- But...

Further Research



Simpson's Paradox !

The difference in midwest and northeast region is significant, but their population is less than a half.

Midwest

Midwest	Inexperienced	Experienced
Black	153	154
White	2366	2179

South

South	Inexperienced	Experienced
Black	577	568
White	2673	2609

Northeast

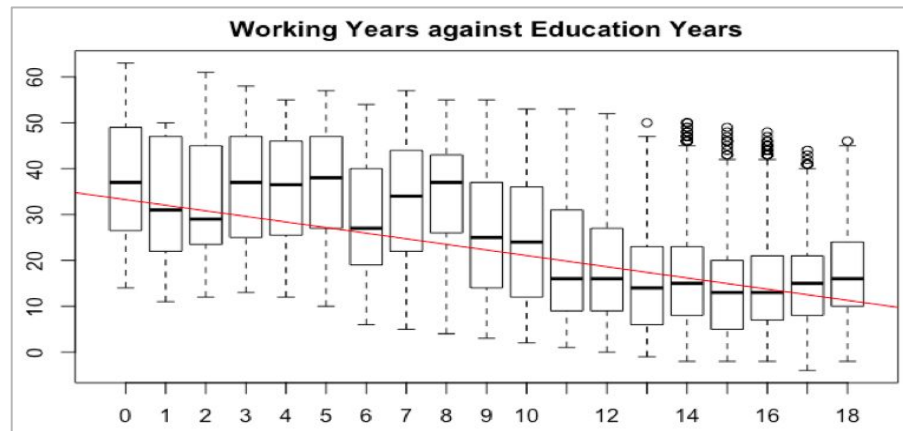
Northeast	Inexperienced	Experienced
Black	134	183
White	2144	2189

West

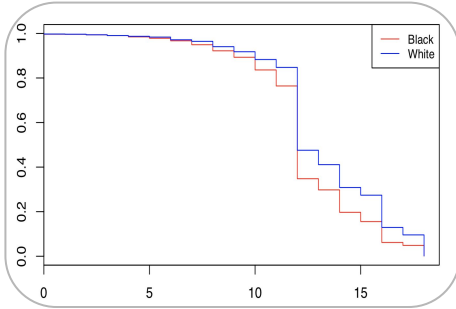
West	Inexperienced	Experienced
Black	90	75
White	2150	1995

Mantel-Haenszel Test

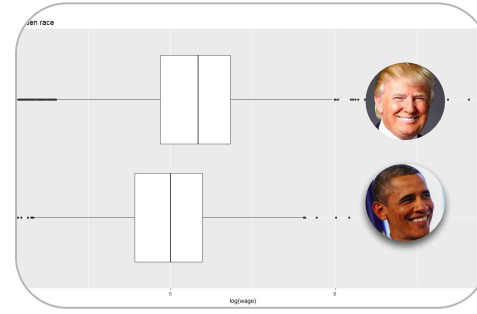
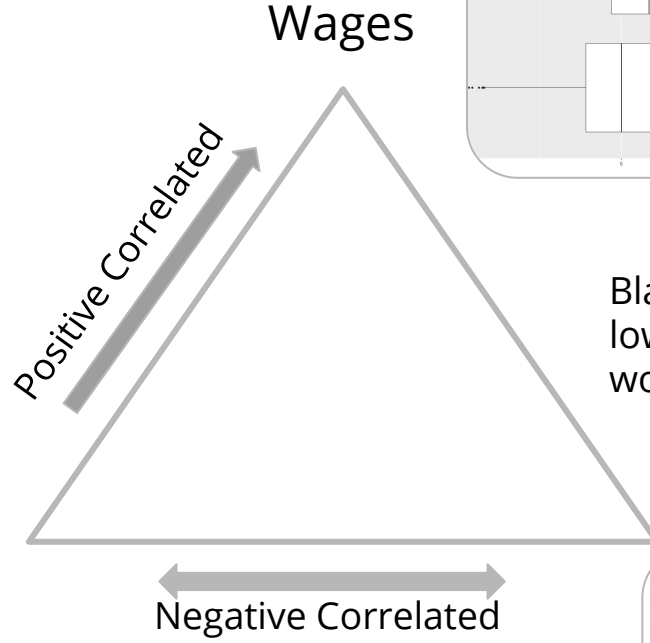
- In M-H test, **p-value=0.0946**, which is different from that of Chi-square test (**p-value=0.1728**). So Black have more working experience.
- Recall the education they received, it leads to the conclusion that Black tend to go to work with lower degree.



Conclusion

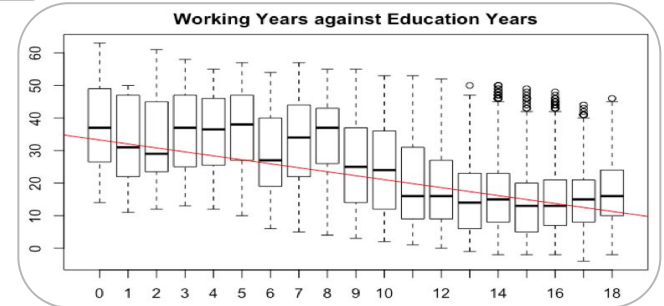


Education Years



Black people tend to have lower wage but more work experience

Work Experience



Thank you!