14. JANUAR 2020

INVESTIGATING SOCIO-ECONOMIC GAPS

IN A HISTORIC US CENSUS DATASET

P. SCHWARZ

Introduction

In public discourse and scientific research, the so-called gender pay gap is currently a reoccurring topic. This is a socio-economic term that describes the difference in hourly gross wage between the genders, usually assumed to be in favor of men. This gap is of different sizes depending on country/culture. Different reasons are brought forward for the existence of the gender pay gap, for example (Blau & Kahn, 2017; OECD, 2020):

- men supposedly work more extra hours which are also higher paid
- women stay often at home to care about their children, which lowers the total work time during their lifetime and might limit career advances due to longer absences from the work life
- Women chose to prefer different kinds of jobs, which are often lower paid
- Women are more defensive in salary negotiations.

This study seeks out to see if there are potentially other types of pay gaps in the population that are not so often topic of public discourse evident in the data.

It is based on a historic data set by the US census bureau from the year 1994. This inherently means that it only covers data of one country and also, by choice one single year in the past. This choice was made for ease of access to the data and to limit the computational capacity needed for this study as resources were very limited.

Firstly, it is the aim to find out of the gender pay gap itself is even evident in the available data set before looking for more. This is the basis for the following research question:

RQ1: Is there a difference in hourly pay in the 1994 US census respondents based on gender?

To answer this research question the following testable hypothesis has been developed:

Hypothesis1:

Among the 1994US census respondents, there is no significant difference between hourly wage for adults based on gender

One determinant for wage is education. The longer someone is educated the higher the wage is statistically (Afxentiou , Kutasovic, 2010) This leads to the question if in there is evidence in the gathered data that speaks for a gender disparity in length of education. As one point of interest it was decided to look at the highest possible level of education captured in the census data. The idea is to see if education could be a component of a possibly existing pay gap in the dataset, leading to the following research question:

RQ2: Is there a difference in gender when it comes to the highest level of education for the 1994 US census respondents, an education gap?

To answer this research question it was decided to focus on the highest educational status in the dataset, which is a PhD. In current times studies show that in the US there is an education gap but in favor of women (DiPrete & Buchmann, 2013). To see if this was the case in 1994, the following testable hypothesis has been developed:

Hypothesis2:

Among the 1994 US census respondents, there is no difference between the amount of adult men and women with a PhD.

Besides gender there might be other attributes that cause pay gaps between respondents. One of these could be the age of the census respondents. It can be expected that older workers generally have more experience in their profession and therefore get paid more. Evidence has for example been found in Italy (Toepfer, 2019) . However, since this study only looks at respondents that report an hourly wage and most high paying jobs are not actually paid by the working hour this might not be a valid argument in this context. To confirm this idea the following research question is posed:

RQ3: Can the respondent age be used as a predictor for hourly salary in the 1994 US census dataset?

This leads to the third hypothesis to be tested:

Hypothesis3:
Among the 1994 US census respondents, there is no statistically significant relationship between respondent age and hourly wage for adults

There are many studies that have race inequalities as a subject, especially in a US context. This naturally sparks the question of potential race pay gaps (Reid, 1998). The question is if there are hourly wage difference visible between races among the respondents of this census dataset:

RQ4: Are there difference in hourly wage among the 1994 census respondents depending on their race?

Leading to the last hypothesis:

Hypothesis4:
Among the 1994 US census respondents, there is no statistically significant in hourly wage difference between races for adults

Method

The research design of this study is experimental, which means hypothesis are constructed to test relationships based on empirical data (Johanesson & Perjons, 2014, pp.40-42).
It has to be mentioned that the dataset is based on a stratified sample by the US census bureau in order to create representative statistics and every row in the dataset therefore is coming with a weight to convey how many people in the total population are represented by each sample. This has been disregarded in this study as the scope of the research questions are only the actual responses to the census.

For all the hypothesis testing methodology following in this section it can be stated that generally it is preferred to conduct a parametric test due to their advantageous statistical strength. Therefore, every time there is a parametric and a non-parametric alternative for testing the hypothesis the satisfaction of the assumptions for the parametric test will be assessed first. (Denis, 2016, p.149)

Hypothesis 1 has with wage per hour a numerical dependent variable and a categorical independent variable with the sex of the respondents. The aim is to look for difference between male and female respondents and their hourly wage in cent. If the underlying assumptions are fulfilled a t-test can be used to compare the mean wages of the groups respondents. The non-parametric alternative is a Wilcoxon rank sum test.

Hypothesis2 is based on two categorical variables sex and education with 2 groups, therefore comparing nominal data. Considering the high number of respondents only a chi square test is an option to test this hypothesis. The two different chi-square tests are the chi-square test for association and the chi-square test for differences. Since there are no expected values provided nor can be deducted, the relevant test type is the chi-square test for association. (Denis, 2016, p.92)

Hypothesis 3 deals with two continuous numerical values and their interaction.
Since the research question is aimed at prediction a linear regression is the primary option if the respective underlying assumptions are satisfied. (Denis, 2016, pp.333-334)

Hypothesis 4 deals with 3 or more groups, which means that the parametric test of choice would be an ANOVA, the non-parametric alternative in case of violated assumptions is the Kruskal

Wallis rank sum test. The test result will demonstrate the potential difference between all pairs of groups. (Denis, 2016, pp.226-228)

The dataset has been filtered after it was downloaded from the University of California Irvine Machine Learning Repository (Dua and Graff, 2019) where it is available under the following url: https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29 . Since it included data from two years the irrelevant year has been filtered out in Excel. A similar filtering was conducted in order to test the second hypothesis. For the testing of hypothesis 1, 3 and 4 only respondent data with an hourly salary was used. Also, only respondents older than 18 have been considered in general.

Results

Hypothesis 1

The first hypothesis deals with the gender age gap. As mentioned in the previous section it was first checked if the assumption for conducting a t-test were satisfied.

These assumptions are (Elrod, cornell.edu, nd):

- The dependent variable, in this case hourly wage is normally distributed.
- The independent variable needs to be bivariate
- The dependent variable is continuous.
- Each dependent variable observation is independent

The observations are independent form each other, the independent variable is bivariate (male/female) and the hourly wage is a continuous variable.

To assess the distribution of the dependent variable a histogram was created,

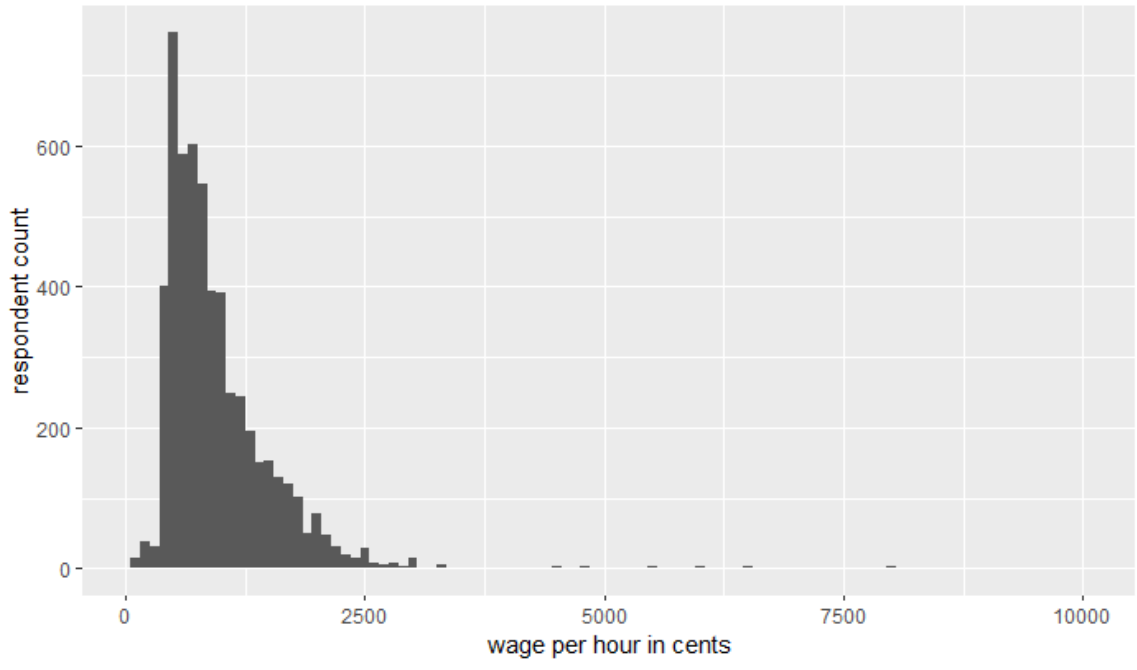


Figure 1: Histogram to assess the distribution of the dependent variable

Dividing the hourly wage by gender gives the following result:

Figure 2: Histogram to assess the distribution of the dependent variable for each sex

After the assessment of the histogram it could be concluded that the normality assumption of the dependent variable was violated as the data for both sexes has a positive skew. Therefore, it has been decided to use the non-parametric alternative, the Wilcoxon rank sum test since the samples are independent.

The following boxplot has been created to demonstrate the medians for both sexes as they are relevant for the Wilcoxon rank sum test.

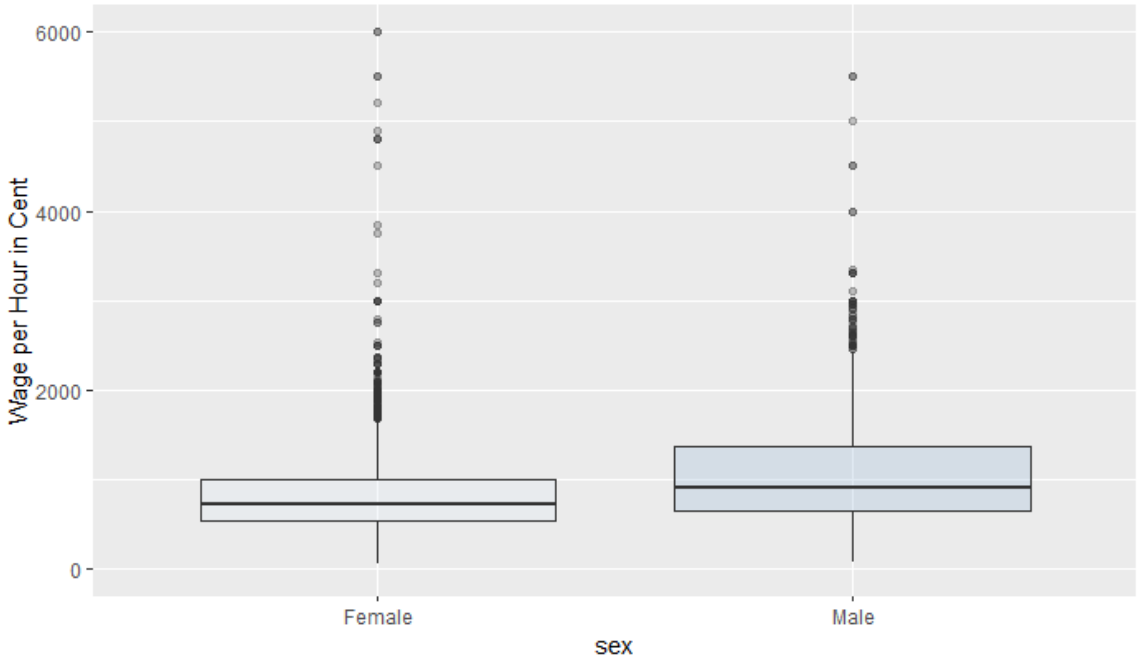


Figure 3: Boxplot with hourly wages for the sexes (21 outliers truncated for scale)

The test provided the following result

Table 1: Wilcoxon Rank Sum Test for wage per hour by sex

Female respondents	Male respondents	W-statistic	p-value
726	918	2850019	0.00000000000000022

The result of the Wilcoxon test is W=2850019, a p-value of < 2.2e-16 with n=1644

This means the Null hypothesis is rejected as p is smaller than 0.05, which indicates there is a significant difference in median for hourly wage based on gender among the adult respondents of the 94 census.

Hypothesis 2

After counting the male and female respondents with and without a PhD based on a table pre-filtered in Excel and loaded into R, the following contingency table was created:

Table 2: Contingency Table for having a PhD and gender

	Female	Male
PhD	173.00	443.00
No PhD	37991.00	33113.00

In this contingency table none out of four has a cell value below 5. Therefore, a chi-square test for associations can be used without Yates correction as the total N of the exceeds 40. (Heukman & Brunner, 2016 and Hoffman, 2019)

The assumptions required for a chi square-test (McHugh, 2013) are:

1. *Contingency table cells are counts of cases*

   This is the case as the cells are filled with counts of census respondents

2. *Mutually exclusive variable categories*

   Respondents cannot be male and female, nor can they be a PhD and not a PhD

3. *Each subject only contributes data to one cell*

   This is the case as each respondent can only be either male or female and only be a PhD or not a PhD but not both

4. *The study groups are independent*

   There is no connection between the respondent groups

This mean all requirements to conduct a chi-squared test for association are satisfied. The test results are summarized in the following table:

Table 3: Pearson's Chi-squared test for association

χ^2	dF	p-value
157.58	df = 1	2.2e-16

The test statistic is 157.58 at 1 degree of freedom with a p-value of <2e-16 for the chi square test for association without Yates correction. This means that the null hypothesis can be rejected at a 0.05 level. Therefore, we can conclude that among the 1994 US census respondents, there is a difference between the amount of adult men and women with a PhD.

Hypothesis 3

A linear regression model with census respondent age as the independent variable and hourly wage as the dependent variable has been created. Linear regressions require several assumptions to be met (Denis, 2016, pp.346-347):

- Linearity
- Independence
- Normality
- Equality of variance

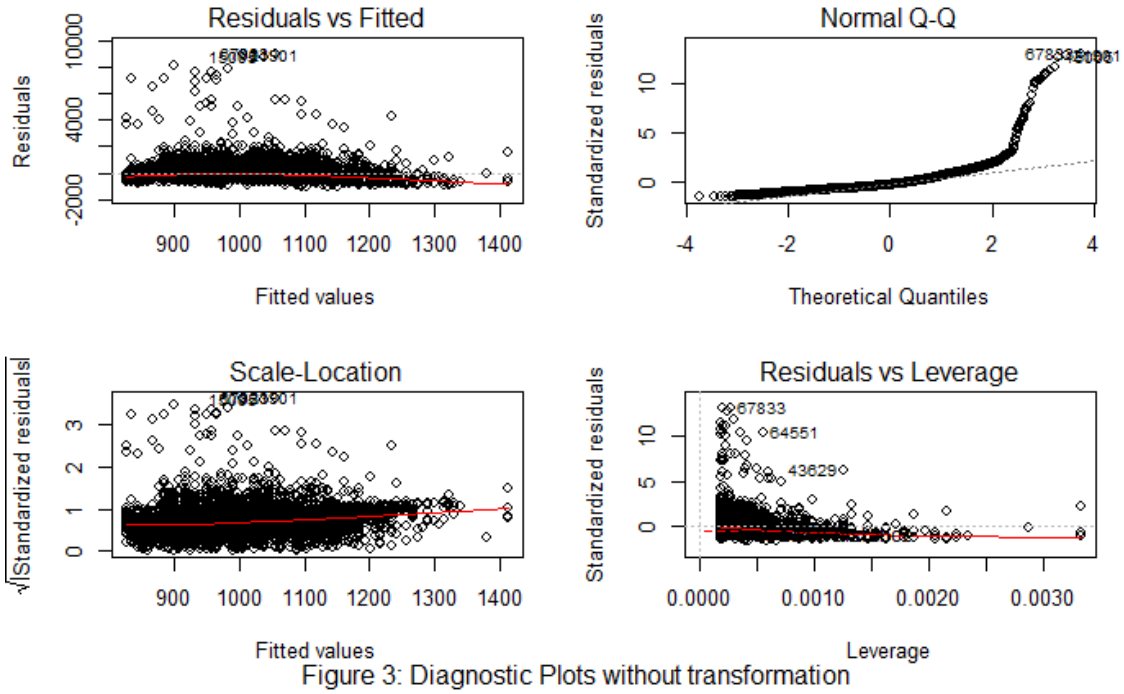


Figure 3: Diagnostic Plots without transformation

Residuals vs fitted barely demonstrates a difference between the cases. The QQ plot tells us that the residuals are diverging dramatically from the line the further one moves to the right along the x-axis. The redline indicate a limited residual spread. In residuals vs leverage the cook distance lines are not visible, which means no influential data points outside the area.
(University of Virginia Data Library Virginia.edu, 2015)

Even though one could argue the central limit theorem does let one assume normality, a log10 transformation on the hourly wages was conducted to fulfill the requirements of the linear regression, even though this is not without problems. Residuals vs fitted show more of a difference between the cases. The QQ plot is much closer to normality except on the tails. For

scale-location the red line is sloping slightly indicating problems with the constant variance assumption. Very few residuals are outside the cook line and therefore influential. It has been decided to not remove these due to a lack of substantive reason to remove them (University of Virginia Data Library Virginia.edu, 2015; Denis, 2016, p.347)

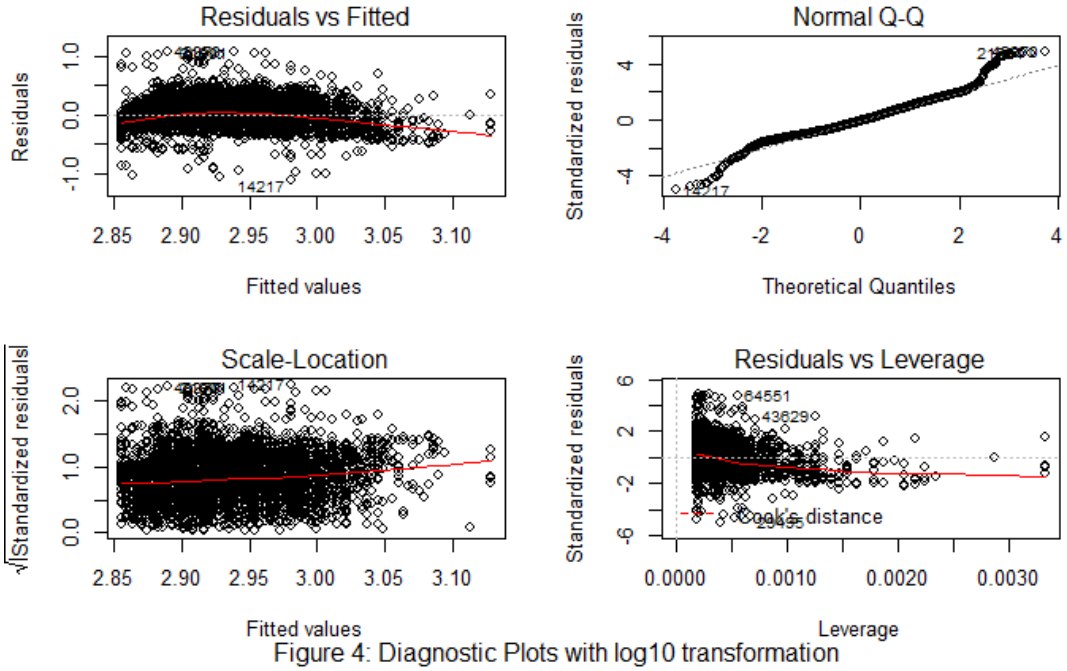


Figure 4: Diagnostic Plots with log10 transformation

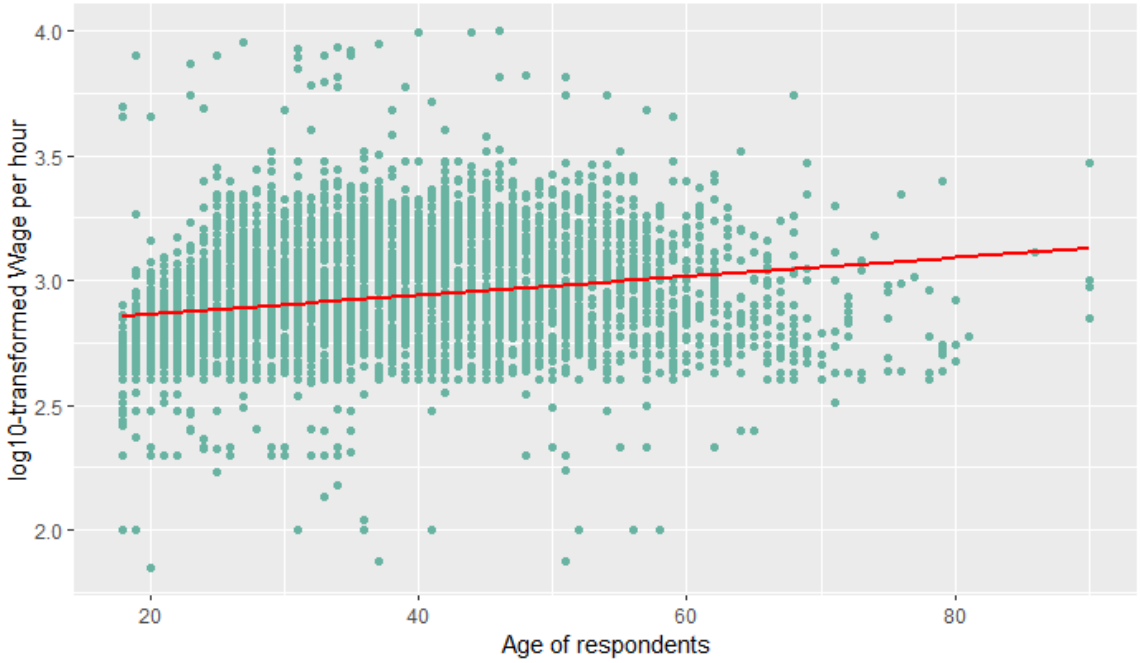


Figure 5: Scatterplot of log10-transformed hourly wage by age with regression line

The graph above shows all data points with the regression line sloping slightly upwards and confirms the linearity assumption.

Table 4: Linear Regression with Age and log transformed Hourly Wage Part 1

	Dependent variable: wage_per_hour
Observations	5,468
R^2	0.022
Adjusted R^2	0.022
Residual Std. Error	686.486 (df = 5466)
F Statistic	124.478*** (df = 1; 5466)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Linear Regression with Age and log transformed Hourly Wage Part 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	680.0329	28.7493	23.65	0.0000
age	8.1433	0.7299	11.16	0.0000

R^2 is 0.022, b is 8.1433, the standard error is b is 0.7299, t is 11.16 p is below 0.01. The F statistic is 124.478 on 1 and 5466 DF. Since the dependent variable has been log-transformed the independent variable has a multiplicative relationship with the dependent variable instead of and additive relationship. (University of Virginia Data Library Virginia.edu, 2018)

The result shows that since the p-value is close lower than 0.05, the Null hypothesis can be rejected. This means that among the 1994 US census respondents, there is a statistically significant relationship between respondent age and hourly wage for adults.

Hypothesis 4

From the previous results it is already known that hourly wage is not normally distributed. Hence hypothesis 4 needs to be tested with a non-parametric test. In this case the Kruskal-Wallis test was conducted. Since the test is based on the medians of the groups they are listed in the next table:

Table 6: Median hourly wage per race

Race	Median Hourly Wage
Amer Indian Aleut or Eskimo	855
Asian or Pacific Islander	903
Black	725
Other	625
White	800

The table below summarizes the test result of the Kruskal-Wallis test for wage per hour grouped by race:

Table 7: Kruskall-Wallis Rank Sum Test

Data:	Hourly Wage and Race
Kruskal-Wallis χ^2	32.177
df	4
p-value	0.00000176

The test statistic is 32.177 with 4 degrees of freedom and the p-value is 0.00000176, which is below 0.05. Consequently, the null hypothesis can be rejected. Among the 1994 US census respondents, there is a statistically significant hourly wage difference between races.

To find out which groups have significantly different median hourly wages compared to each other the following post-hoc comparison table was computed:

Table 8: Multiple comparison test after Kruskal-Wallis

p.value: 0.05 Comparisons	obs.dif	critical.dif	difference
Amer Indian Aleut or Eskimo- Asian or Pacific Islander	49.11499	692.1382	FALSE
Amer Indian Aleut or Eskimo- Black	469.00216	583.9178	FALSE
Amer Indian Aleut or Eskimo- Other	637.32758	873.7615	FALSE
Amer Indian Aleut or Eskimo- White	115.80982	557.6859	FALSE
Asian or Pacific Islander- Black	518.11715	454.3060	TRUE
Asian or Pacific Islander- Other	686.44257	793.0278	FALSE
Asian or Pacific Islander- White	164.92482	420.0565	FALSE
Black- Other	168.32542	700.5696	FALSE
Black- White	353.19234	195.8373	TRUE
Other- White	521.51776	678.8603	FALSE

From the post-hoc test it can be concluded that the race differences in hour wages stem from two comparisons. Firstly, the group Asian or Pacific Islander has a significantly higher median hourly wage than the group Black. Secondly, the group Black has a significantly lower hourly wage than the group White. All other group comparisons were insignificant at the 0.05 level.

The following boxplot visualizes the wage differences between the five groups:

Figure 6: Boxplot of hourly wage and race of respdondents

In the boxplot it can be observed that that the differences between the groups Black and Asian or Pacific Island and Black and White are the most obvious ones.

Discussion

The results indicate that there are many socio-economic gaps between groups in the 1994 US census dataset. Men have a higher hourly wage, as was initially expected. There are also hourly wage gaps between some races, but not all of them, which is somewhat interesting, but could for example be partly geographically explained, but that is a question for further research. There is a relationship between age and hourly salary, showing that older people earn more. As mentioned before this could be explained by experience being paid for. But the scatterplot shows that in very high ages the hourly wages seem to go down. An option for further research could be to cut off all data above the retirement age to get a better representation. The study also showed an education gap between men and women when it comes to PhD level. It would be interesting to see if this is still the case today and how this influences the gender pay gap.

To advance the study further multivariate test could be conducted and several control variables could be used as this was out of scope for this study and it would help significantly in investigating cause and effect relationships in the data. Also, many columns in the dataset have been unused because they were out of scope, that could provide more valuable insights. Another option would be a year over comparison between different census year datasets.

References

Afxentiou, D., Kutasovic, P., 2010. Does College Education Pay? Evidence from the NLSY-79 Data. Contemporary Issues in Education Research; Littleton 3, 119–126.

Blau, F.D., Kahn, L.M., 2017. The Gender Wage Gap: Extent, Trends, and Explanations. Journal of Economic Literature; Nashville 55, 789–865.

Denis, D. J. (2016). Applied univariate, bivariate, and multivariate statistics. Hoboken, New Jersey: John Wiley & Sons, Inc.

DiPrete, T.A., Buchmann, C., 2013. The Rise of Women: The Growing Gender Gap in Education and What it Means for American Schools. Russell Sage Foundation, New York, UNITED STATES.

Dua, D., Graff, C., 2017. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.

Earnings and wages - Gender wage gap - OECD Data [WWW Document], 2020 . the OECD. URL http://data.oecd.org/earnwage/gender-wage-gap.htm (accessed 1.14.20). http://dx.doi.org.ezproxy.its.uu.se/10.1257/jel.20160995

Elrod, Assumptions for the t-test [WWW Document], n.d. URL http://www.csic.cornell.edu/Elrod/t-test/t-test-assumptions.html (accessed 1.14.20).

Interpreting Log Transformations in a Linear Model | University of Virginia Library Research Data Services + Sciences [WWW Document], 2018 URL https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/ (accessed 1.14.20).

Johannesson, P., Perjons, E., 2014. An Introduction to Design Science. https://doi.org/10.1007/978-3-319-10632-8

Julien I.E. Hoffman, in Basic Biostatistics for Medical and Biomedical Practitioners (Second Edition), 2019

McHugh M. L. 2013. The chi-square test of independence. Biochemia medica, 23(2), 143–149. doi:10.11613/bm.2013.018

Reid, L.L., 1998. Devaluing Women and Minorities: The Effects of Race/Ethnic and Sex Composition of Occupations on Wage Levels. Work and Occupations 25, 511–536. https://doi.org/10.1177/0730888498025004005

Timothy Beukelman, Hermine I. Brunner, in Textbook of Pediatric Rheumatology (Seventh Edition), 2016

Töpfer, M., 2019 The Age Pay Gap and Labour Market Heterogeneity: A New Empirical Approach Using Data for Italy. LABOUR n/a. https://doi.org/10.1111/labr.12161

Understanding Diagnostic Plots for Linear Regression Analysis | University of Virginia Library Research Data Services + Sciences [WWW Document], 2015 URL https://data.library.virginia.edu/diagnostic-plots/ (accessed 1.14.20).

Appendix

R Code

Patrick Schwarz

13 Januar 2020

INITIAL LOADING OF PACKAGES AND THE DATASET

```
library(Rcmdr) #this chunk start RCommander

library(ggplot2) #this cunks starts ggplot2
library(hrbrthemes)

library(pgirmess)
library(plyr)
library(xtable)

all_data <-  read.csv("adults 94.csv",
                 header = TRUE,
                 sep= ";")

phds <-  read.csv("phds.csv",
                 header = TRUE,
                 sep= ";")

hourly_paid_adults <-all_data[all_data$wage_per_hour >0,]
```

HYPOTHESIS 1

```
ggplot(hourly_paid_adults, aes(x=wage_per_hour)) + geom_histogram(binwidth
= 100) +
  labs(caption= "Figure 1: Histogram to assess the distribution of the dep
endent variable") +
  xlab("wage per hour in cents") +
  ylab("respondent count") +
  theme(plot.caption = element_text(hjust=0.5, size=rel(1)))
```

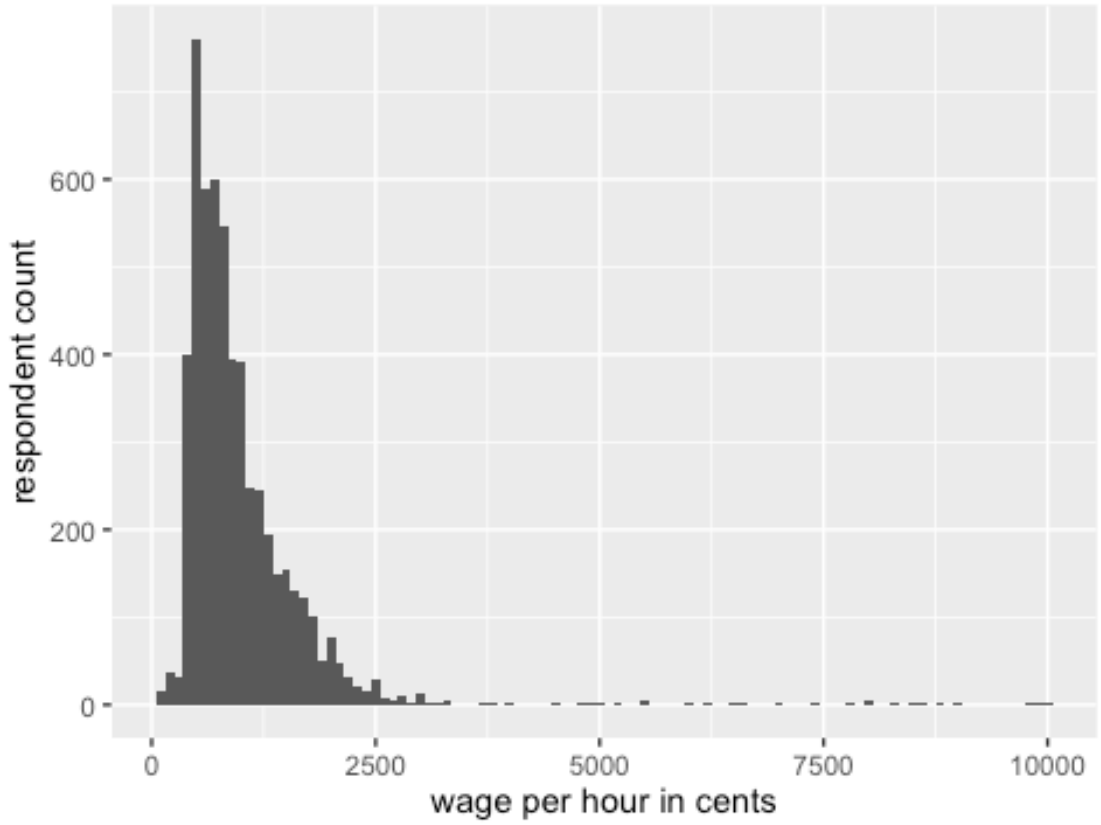
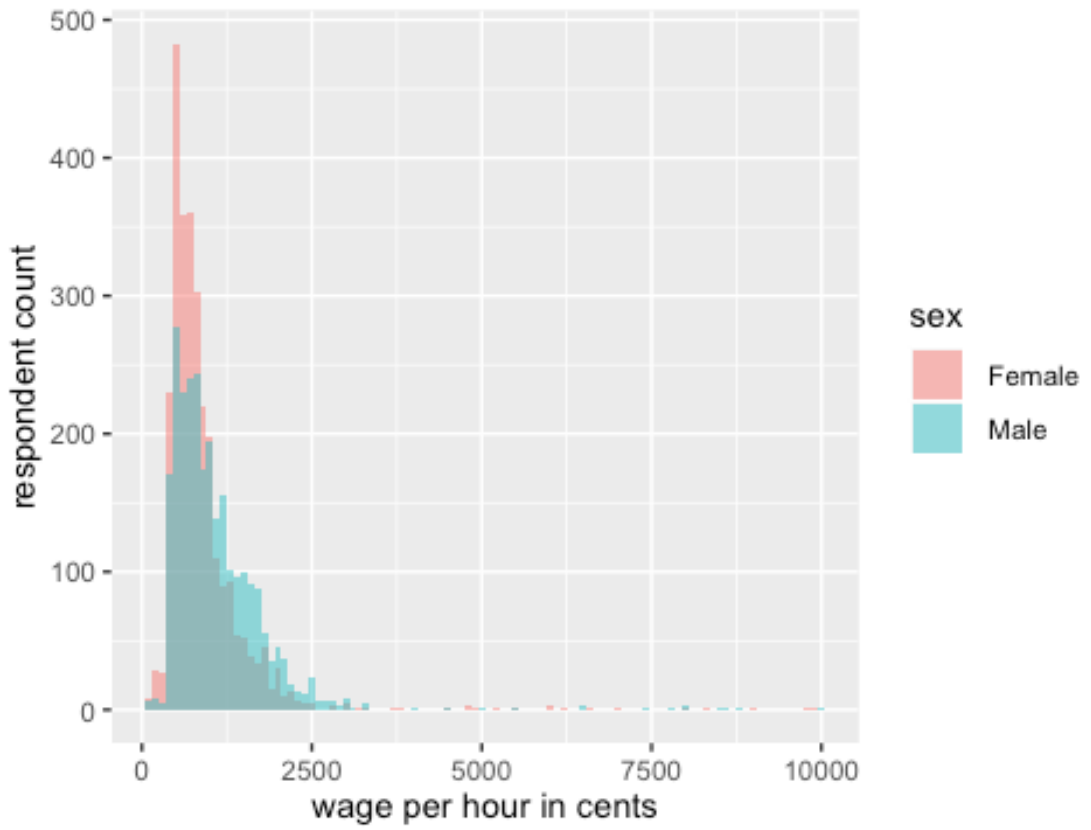


Figure 1: Histogram to assess the distribution of the dependent variab

```
ggplot(hourly_paid_adults, aes(x=wage_per_hour, fill=sex)) +
    geom_histogram(binwidth=100, alpha=.5, position="identity") +
    xlab("wage per hour in cents") +
    ylab("respondent count") +
    labs(caption= "Figure 2: Histogram to assess the distribution of the d
ependent variable for each sex") +
    theme(plot.caption = element_text(hjust=0.5, size=rel(1)))
```



Histogram to assess the distribution of the dependent variable for each se

```
ggplot(hourly_paid_adults, aes(x=sex, y=wage_per_hour, fill=sex)) +
    geom_boxplot(alpha=0.3) +
    theme(legend.position="none") +
    scale_fill_brewer(palette="BuPu") +
    ylim(0, 6000) +
    ylab("Wage per Hour in Cent")+
    labs(caption= "Figure 3: Boxplot with hourly wages for the sexes (21 o
utliers truncated for scale)") +
    theme(plot.caption = element_text(hjust=0.5, size=rel(1)))
```

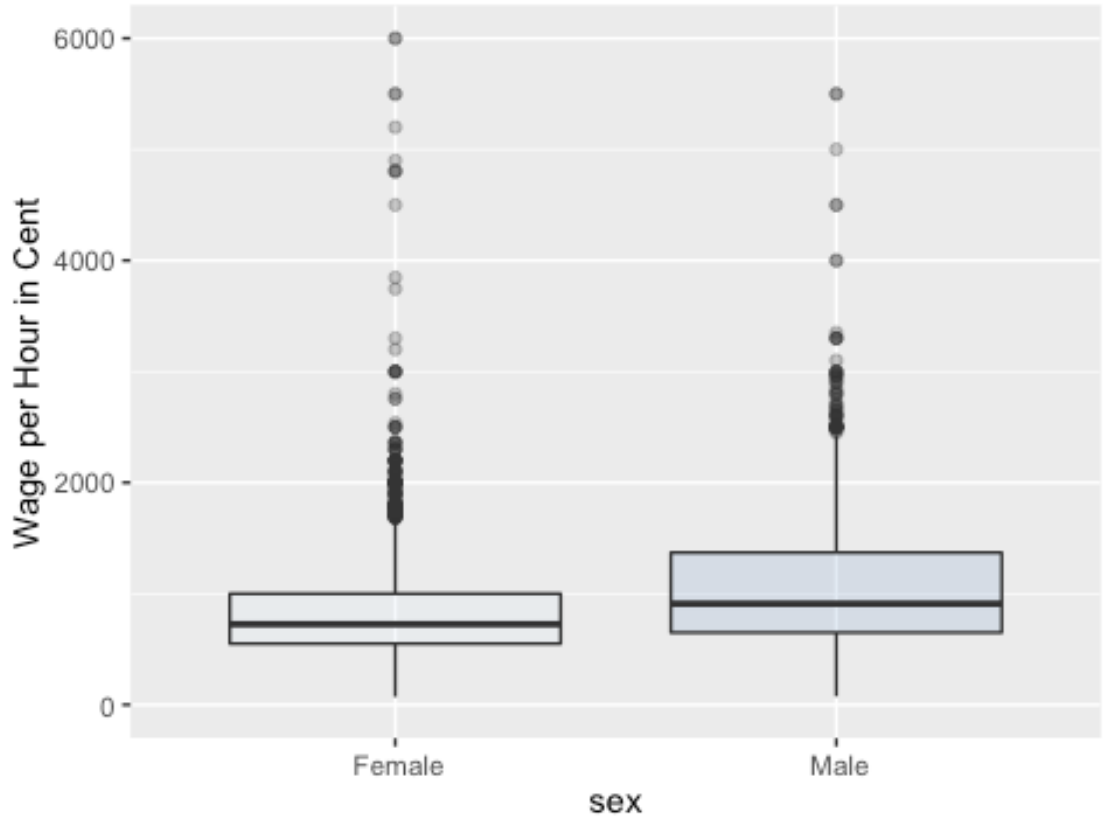


Figure 3: Boxplot with hourly wages for the sexes (21 outliers truncated fo

```
with(hourly_paid_adults, tapply(wage_per_hour, sex, median, na.rm=TRUE))
```

```
##  Female    Male
##     726     918
```

```
wcoxdifference_test <- wilcox.test(wage_per_hour ~ sex, alternative="two.s
ided", data=hourly_paid_adults)
wcoxdifference_test
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  wage_per_hour by sex
## W = 2850019, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

HYPOTHESIS 2

```
count(phds$sex)
```

```
##         x freq
## 1  Female  173
## 2    Male  443
```

```
count(all_data$sex)
```

```
##         x  freq
## 1  Female 38164
## 2    Male 33556
```

```r
count(all_data$sex)-count(phds$sex)
```

```
##    x  freq
## 1 NA 37991
## 2 NA 33113
```

```r
.Table <- matrix(c(173,443,37991,33113), 2, 2, byrow=TRUE)
dimnames(.Table) <- list("education"=c("PhD", "No PhD"), "sex"=c("Female",
"Male"))
.Table  # Counts
```

```
##           sex
## education Female  Male
##    PhD       173   443
##    No PhD  37991 33113
```

```r
.Test <- chisq.test(.Table, correct=FALSE)
.Test
```

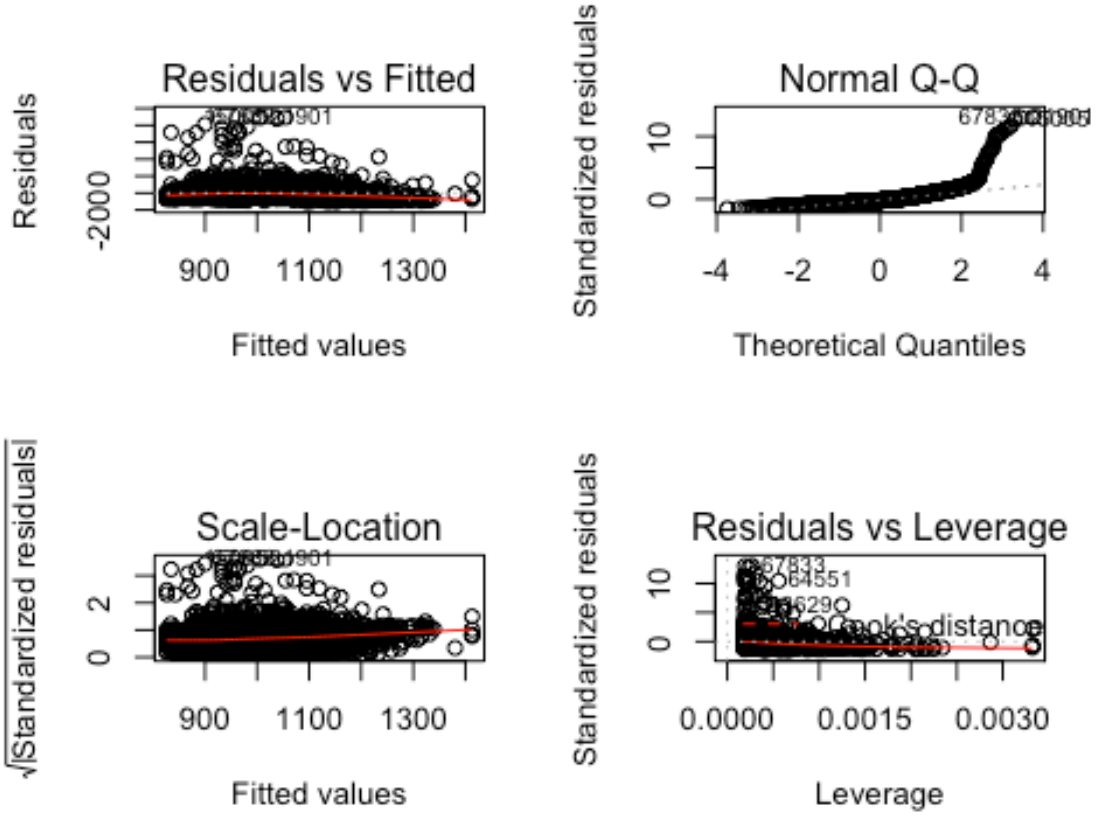
```
##
##  Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 157.58, df = 1, p-value < 2.2e-16
```

```r
xtable(.Table)
```

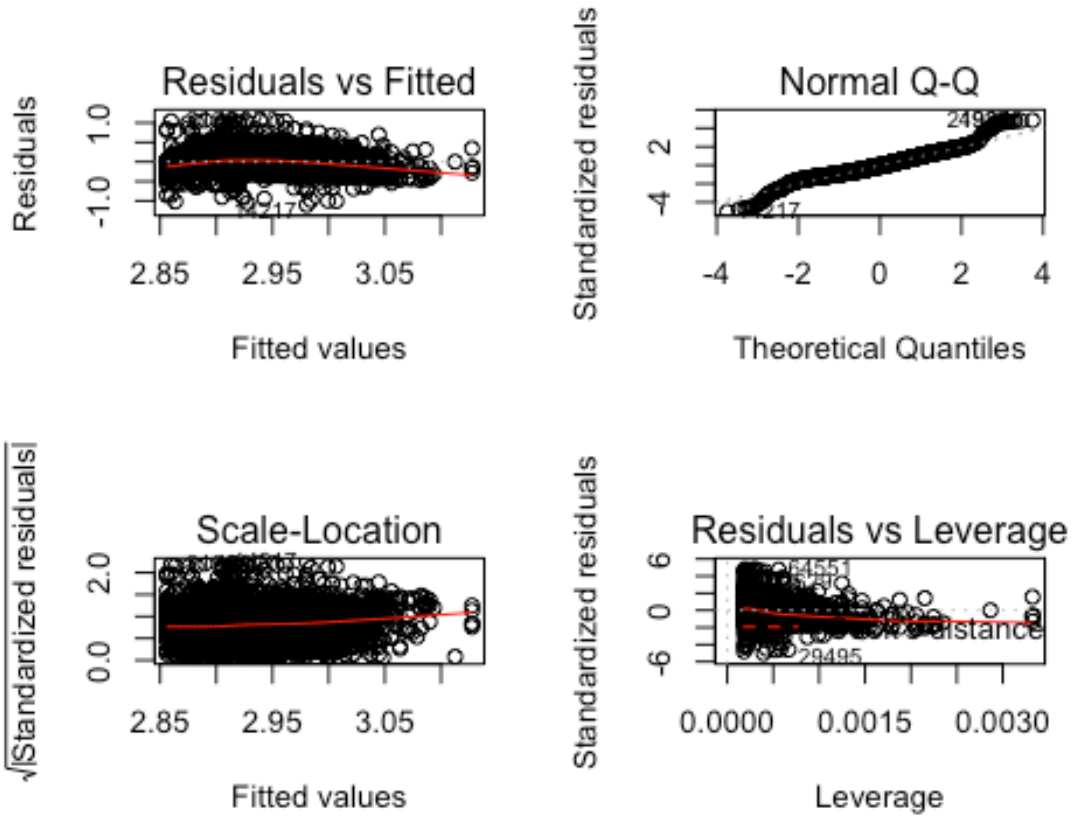
```
## % latex table generated in R 3.6.1 by xtable 1.8-4 package
## % Tue Jan 14 21:07:27 2020
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & Female & Male \\
##   \hline
## PhD & 173.00 & 443.00 \\
##   No PhD & 37991.00 & 33113.00 \\
##    \hline
## \end{tabular}
## \end{table}
```

HYPOTHESIS 3

```r
regression.model <- lm((hourly_paid_adults$wage_per_hour)~(hourly_paid_adu
lts$age_),data=hourly_paid_adults)
par(mfrow=c(2,2))
plot(regression.model)
mtext("Figure 3: Diagnostic Plots without transformation", side=3, outer=T
RUE, line=-27)
```




```
log10regression.model <- lm(log10(hourly_paid_adults$wage_per_hour)~(hourl
y_paid_adults$age_),data=hourly_paid_adults)
par(mfrow=c(2,2))
plot(log10regression.model)
mtext("Figure 4: Diagnostic Plots with log10 transformation", side=3, oute
r=TRUE, line=-27)
```




```
summary(regression.model)

##
## Call:
## lm(formula = (hourly_paid_adults$wage_per_hour) ~ (hourly_paid_adults$a
ge_),
##     data = hourly_paid_adults)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1052.3  -368.6  -159.2   188.4  8944.4
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               680.0329    28.7493   23.65   <2e-16 ***
## hourly_paid_adults$age_     8.1433     0.7299   11.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 686.5 on 5466 degrees of freedom
## Multiple R-squared:  0.02227,    Adjusted R-squared:  0.02209
## F-statistic: 124.5 on 1 and 5466 DF,  p-value: < 2.2e-16

 ggplot(hourly_paid_adults, aes(x=hourly_paid_adults$age, y=log10(hourly_p
aid_adults$wage_per_hour))) +
  geom_point( color="#69b3a2") +
  geom_smooth(method=lm , color="red", se=FALSE) +
  ylab("log10-transformed Wage per hour") +
  xlab("Age of respondents") +
```

```
  labs(caption= 'Figure 5: Scatterplot of log10-transformed hourly wage by
age with regression line') +
  theme(plot.caption = element_text(hjust=0.5, size=rel(1)))
```

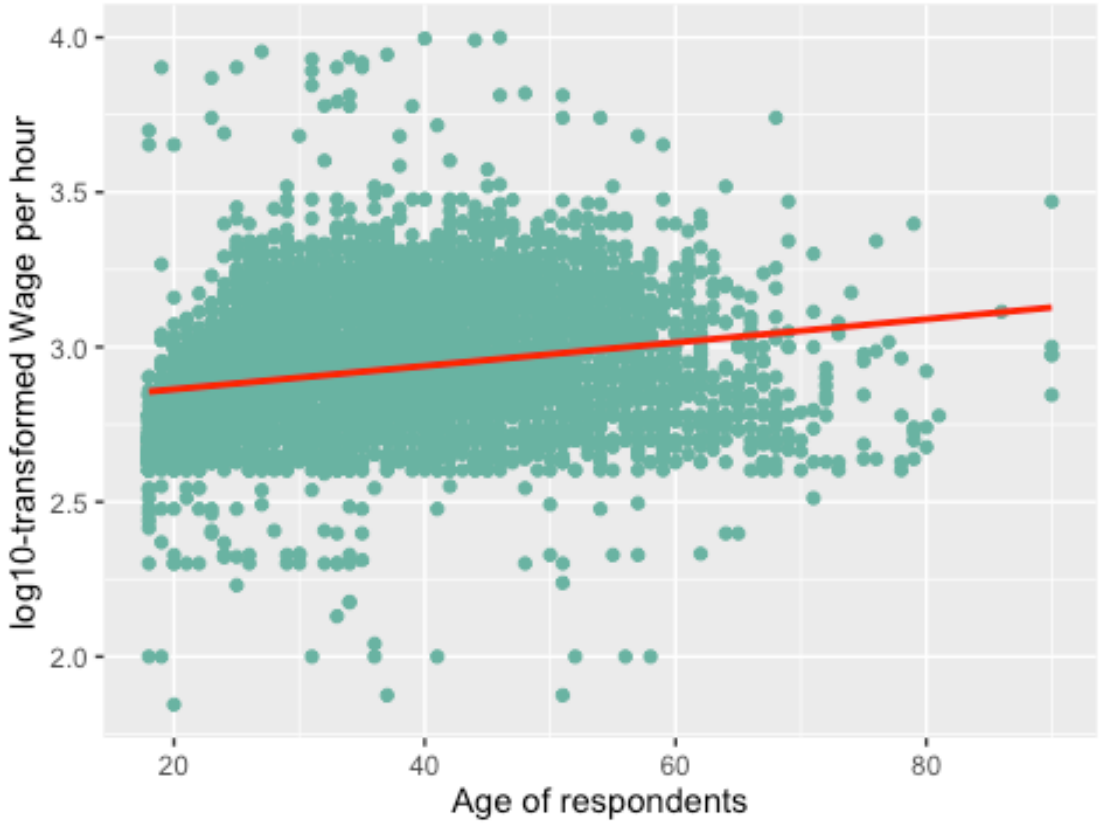


Figure 5: Scatterplot of log10-transformed hourly wage by age with regress

HYPOTHESIS 4

```
ktest <-kruskal.test(hourly_paid_adults$wage_per_hour , hourly_paid_adults
$race)
ktest

##
##  Kruskal-Wallis rank sum test
##
## data:  hourly_paid_adults$wage_per_hour and hourly_paid_adults$race
## Kruskal-Wallis chi-squared = 32.177, df = 4, p-value = 1.76e-06

with(hourly_paid_adults, tapply(wage_per_hour, race, median, na.rm=TRUE))

##  Amer Indian Aleut or Eskimo    Asian or Pacific Islander
##                          855                          903
##                        Black                        Other
##                          725                          625
##                        White
##                          800

wage_race <- kruskalmc(hourly_paid_adults$wage_per_hour, hourly_paid_adult
s$race, probs = 0.05, cont=NULL)

wage_race
```

```
## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
##                                                          obs.dif
##  Amer Indian Aleut or Eskimo- Asian or Pacific Islander  49.11499
##  Amer Indian Aleut or Eskimo- Black                     469.00216
##  Amer Indian Aleut or Eskimo- Other                     637.32758
##  Amer Indian Aleut or Eskimo- White                     115.80982
##  Asian or Pacific Islander- Black                       518.11715
##  Asian or Pacific Islander- Other                       686.44257
##  Asian or Pacific Islander- White                       164.92482
##  Black- Other                                           168.32542
##  Black- White                                           353.19234
##  Other- White                                           521.51776
##                                                       critical.dif
##  Amer Indian Aleut or Eskimo- Asian or Pacific Islander     692.1382
##  Amer Indian Aleut or Eskimo- Black                         583.9178
##  Amer Indian Aleut or Eskimo- Other                         873.7615
##  Amer Indian Aleut or Eskimo- White                         557.6859
##  Asian or Pacific Islander- Black                           454.3060
##  Asian or Pacific Islander- Other                           793.0278
##  Asian or Pacific Islander- White                           420.0565
##  Black- Other                                               700.5696
##  Black- White                                               195.8373
##  Other- White                                               678.8603
##                                                       difference
##  Amer Indian Aleut or Eskimo- Asian or Pacific Islander      FALSE
##  Amer Indian Aleut or Eskimo- Black                          FALSE
##  Amer Indian Aleut or Eskimo- Other                          FALSE
##  Amer Indian Aleut or Eskimo- White                          FALSE
##  Asian or Pacific Islander- Black                             TRUE
##  Asian or Pacific Islander- Other                            FALSE
##  Asian or Pacific Islander- White                            FALSE
##  Black- Other                                                FALSE
##  Black- White                                                 TRUE
##  Other- White                                                FALSE
```

```
ggplot(hourly_paid_adults, aes(x=race, y=wage_per_hour, fill=race)) +
    geom_boxplot(alpha=0.3, outlier.shape=NA) +
    theme(legend.position="none") +
    scale_fill_brewer(palette="BuPu") +
    ylim(0, 2500) +
    coord_flip()+
    ylab("Hourly Wage in Cents")+
    labs(caption= "Figure 6: Boxplot of hourly wage and race of respdonden
ts") +
    theme(plot.caption = element_text(hjust=0, size=rel(1)))
```

Figure 6: Boxplot of hourly wage and race of r