# Analyzing Discriminatory Trends Using Census Data

Sam Arnts / Group 2

12/7/2021

## Abstract

While discrimination remains in most aspects of society, one context where the effects of discrimination are pronounced and easily identifiable is the workforce. The purpose of this report is to look at possible discriminatory trends within the workforce on the basis of age, gender, country of origin, and race. We ask questions regarding gender and its effects on education level and working age, race and its effect on income and weekly hours worked, education level's effect on hours worked, and national origin's effect on income. In general, we find that minorities or people whose national origin isn't the United States often have trouble achieving lucrative professional positions. Also, we find that women and men on average achieve the same level of education, but the average age of women in the workforce is significantly lower than men, indicating that women on average exit the workforce sooner than men.

## Introduction

The data set which we will be analyzing is from UCI's Machine Learning Repository and is named the *Census Income Data Set*. This data set represents Census data from 1994, and has 32,561 observations with 15 variables. The variables we are concerned about are education level (represented numerically), age, race, average hours worked per week, country of origin, and whether or not an individual makes more than 50,000 in annual income. To better understand these trends, I've devised six questions regarding this data with corresponding hypotheses that we can test using classical statistical methods.

1. *Is being male associated with having a higher level of education as compared to the general population?* I hypothesize that men will on average achieve higher levels of education, as men do not feel the societal pressure women might feel to start a family instead of focusing on pursuing a good education.
2. *Is the proportion of non-white workers who make more than 50k less than the proportion of the average population who makes more than 50k?* I propose that the proportion of non-white workers who make more than 50k is less than the proportion of the general population who makes 50k, as persisting discrimination in the workplace makes it less likely for minorities to either be promoted or hired to lucrative positions in the first place.
3. *Does the median age of workers differ based on gender?* My prediction is that the average age of women in the workforce is less than the average age of men, as women might be encouraged or forced to leave the workforce earlier than men to start a family.
4. *Are the average weekly hours worked for White, Black, Native American, and Asian-Pacific workers all the same?* I hypothesize that at least one of these four groups works a substantially different amount than the others, and would also hypothesize that minorities end up working longer hours than white workers, possibly to make up for lower wages that they are paid due to discriminatory practices.
5. *Is there a linear correlation between education level and hours worked, and can we predict hours worked through linear regression?* I predict that a higher education level is correlated with more hours worked, as a higher education level seemingly makes one more employable.
6. *Are being from the United States and making more than 50k a year associated?* I will test the hypothesis that being from the United States and making more than 50k a year is associated, as people born in the U.S might possess some inherent advantages such as speaking English as their first language.

## Methodologies

**Inquiry 1:** One-sided one-sample t-test

Determines whether there is a significant difference between the means of two groups. In this context, we are looking at the mean level of education of men as compared to the mean level of education for all persons. This test follows a t-distribution with a t-test statistic and n-1 degrees of freedom.

**Inquiry 2:** One-sided one-sample z-test for proportions

Allows us to compare two different proportions and whether or not they are significantly different. Here we will be looking at the proportion of non-white workers who make more than 50k in income versus the proportion of the general population who makes more than 50k in income. The z-test follows a normal distribution with a z-test statistic.

**Inquiry 3:** Wilcoxon Rank Sum Test

Used to compare the median values of two populations, which in our context is the median age of male workers versus the median age of female workers. Wilcoxon only requires the two populations to share the same general shape and not be normally distributed.

**Inquiry 4:** Analysis of Variance (ANOVA)

Compares the sample means of more than two groups, and allows us to identify whether at least one group's mean is different than the others. In our study, we will look at whether White, Black, Asian-Pacific, and American-Indian workers all have an equal median hours worked. The populations must be normally distributed, and use an F-distribution with k-1 and n-k degrees of freedom.

**Inquiry 5:** Spearman's Rank Correlation Coefficient

Is a measure of correlation between two variables x and y, with a high correlation resulting in a coefficient close to 1, and a low correlation with a coefficient close to 0. In our study, we measure the correlation between education level and hours worked.

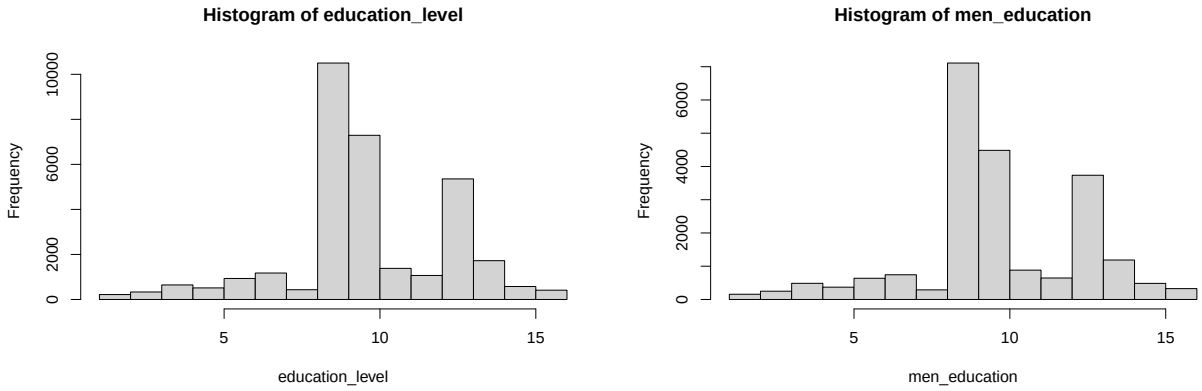
**Inquiry 6:** $\chi^2$ Test of Independence

Tests whether two variables are independent or associated. In our study, we will be testing whether national origin is associated with making more than 50k in annual income. Using a contingency table, the test follows a $\chi^2$ distribution with (r-1) and (c-1) degrees of freedom, where r and c represent the number of rows and columns in the table.

## Inquiries

**Inquiry 1:** Is being male associated with having a higher level of education as compared to the general population?

To conduct this test, we can use the continuous variable "edunum", which is a measure of how many years of education an individual has completed. The lowest number, 1, represents completion of preschool, and the highest number, 16, represents completing a doctorate program. The mean for the entire sample is 10.08068, which indicates some college is completed, while the average for men was 10.10289. While by inspection these means seem almost identical, as well as the two separate distributions, I wondered if the very large sample size will lead to the men's education average being significantly higher than the general population.




```
##
##  One Sample t-test
```

```
##
## data:  men_education
## t = 1.2314, df = 21789, p-value = 0.1091
## alternative hypothesis: true mean is greater than 10.08068
## 95 percent confidence interval:
##  10.07322      Inf
## sample estimates:
## mean of x
##  10.10289
```

$H_0 : \mu \leq 10.081$

$H_1 : \mu > 10.081$

$\alpha = 0.05$

$t = 1.2314$

$p = 0.1091$

When comparing the average population education level to the average education level of men, the men have on average a higher education level. However, after conducting a one-sided t-test, because the p-value $> 0.05$, we fail to reject the null hypothesis, and we have insufficient evidence to support the claim that men have a higher education level as compared to the general population.

**Inquiry 2:** Is the proportion of non-white workers who make more than 50k less than the proportion of the average population who makes more than 50k?

To test this inquiry, we will use the variable "race" to determine whether the subject should be included in the non-white category, as well as the variable "fiftyk", which is a categorical variable that expresses whether an individual makes more than 50k or less than or equal to 50k. The claim we will be testing is that the proportion of non-white individuals who make more than 50k (0.153) is less than the the proportion of people who make more than 50k from the average population (0.241), with both of these numbers being determined from our data set.

```
##
##  Exact binomial test
##
## data:  nrow(minority_50k) and nrow(minority)
## number of successes = 724, number of trials = 4744, p-value < 2.2e-16
## alternative hypothesis: true probability of success is less than 0.2408096
## 95 percent confidence interval:
##  0.0000000 0.1614653
## sample estimates:
## probability of success
##              0.1526138
```

$H_0 : p \geq 0.2408$

$H_1 : p < 0.2408$

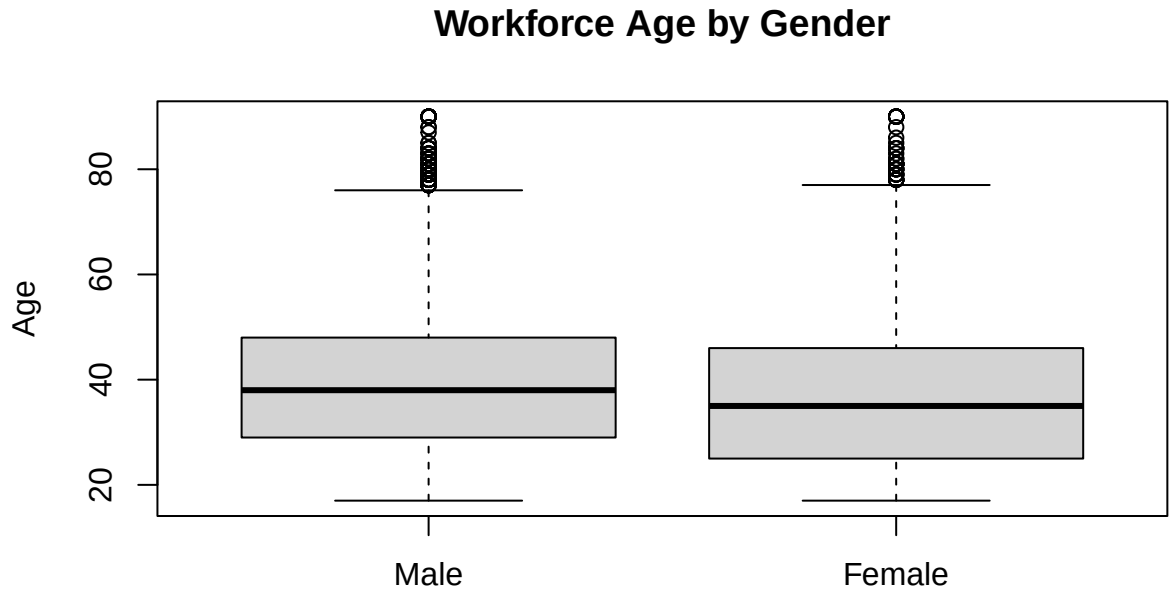
$\alpha = 0.05$

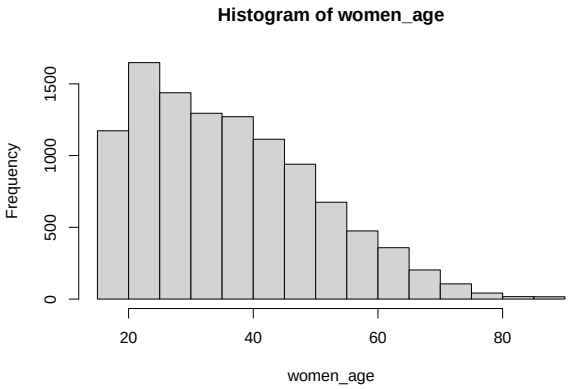
$p < 2.2e - 16$

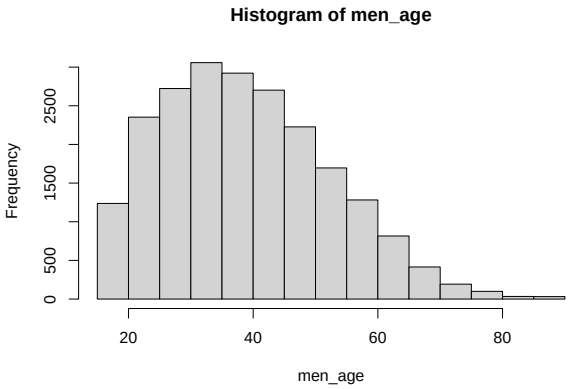
Using a z-test for proportions to test our inquiry, we can see that $p < 0.05$, so we can reject the null hypothesis, and we have significant evidence to conclude that the proportion of non-white workers who make more than 50k is less than the proportion of the average population who makes more than 50k.

**Inquiry 3:** Does the median age of workers differ based on gender?

For this question we will use the variables "gender" and "age" to determine whether there is a significant difference between the median age of men and women in the workforce. After creating a boxplot consisting of these two variables, we can see that men have a marginally higher workforce age. To test our hypothesis, we'll use the nonparametric Wilcoxon rank sum test, which only requires the two populations to have the same general shape, something we can see is true about our samples through plotting histograms.




```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  men_age and women_age
## W = 131800257, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

$H_0$ : The median workforce age for men is the same as the median workforce age for women.

$H_1$ : The median workforce age for men is greater than the median workforce age for women.

$\alpha = 0.05$

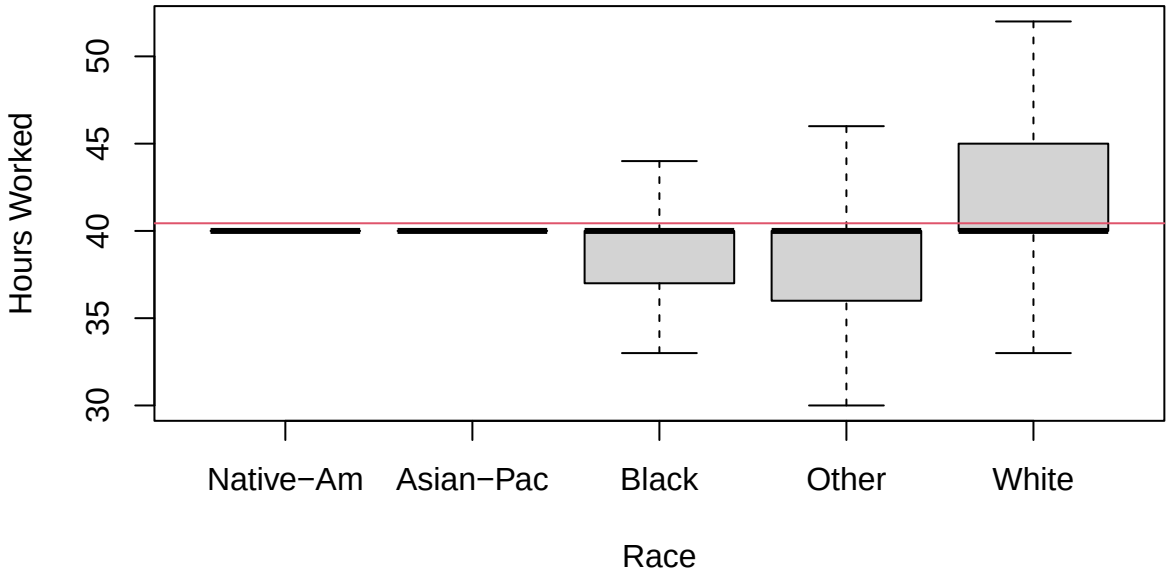
$W = 131800257$

$p = 2.2e - 16$

After conducting our Wilcoxon rank sum test, we see our p-value $< 0.05$, so we can reject our null hypothesis, and we have significant evidence to conclude that the median workforce age differs by gender.

**Inquiry 4:** Are the average weekly hours worked for White, Black, Native American, and Asian-Pacific workers all the same?

In this test, we will be looking at the categorical variable "race" and the continuous variable "hours" which represents the average weekly hours worked of an individual. Creating an ANOVA table will allow us to determine whether each racial group works the same average amount of hours, or whether at least one group works a different amount of hours than another.




```
## Mean White hrs Worked 40.68943

## Mean Black hrs Worked 38.42286

## Mean Asian-Pacific hrs Worked 40.12705

## Mean Native-American hrs Worked 40.01613

## Analysis of Variance Table
##
## Response: hours
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## factor(race)      4   14855  3713.7  24.429 < 2.2e-16 ***
## Residuals     32556 4949210   152.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

$H_1$ : One of the above equalities does not hold

$\alpha = 0.05$

$F_{obs} = 24.408$

$p < 2.2e - 16$

Because our p-value $< 0.05$, our ANOVA allows us to reject the null hypothesis, and we have significant evidence to conclude that at least one of the groups means is statistically different from one of the other groups. We can now further explore our results using a Bonferroni multiple comparison procedure with an overall familywise error rate of $\alpha_{FWE} = 0.05$ to see exactly which group means differ from the others.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  hours and factor(race)
##
##                       Amer-Indian-Eskimo  Asian-Pac-Islander  Black  Other
##  Asian-Pac-Islander 1.0000              -                   -      -
##  Black              0.3001              0.0011              -      -
##  Other              1.0000              1.0000              1.0000 -
##  White              1.0000              1.0000              <2e-16 1.0000
##
## P value adjustment method: bonferroni
```

The results of our Bonferroni multiple comparison test indicates that there is a significant difference between the mean weekly-working hours of Black and White workers as well as Black and Asian-Pacific-Isalnder workers, as comparisons between those groups results in a p-value $< 0.05$.

**Inquiry 5:** Is there a linear correlation between education level and hours worked, and can we predict hours worked through linear regression?

This inquiry can be tested using the Spearman's rank correlation test, which will show us whether any sort of correlation exists between the continuous variable "edunum" and continuous variable "hours". If the Spearman rank correlation coefficient is high (above .85), it will likely indicate that these variables can be used to create a linear regression which will successfully predict hours worked given education level.

```
##
##  Spearman's rank correlation rho
##
## data:  education_level and hours
## S = 4.7915e+12, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1672151
```

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$\alpha = 0.05$

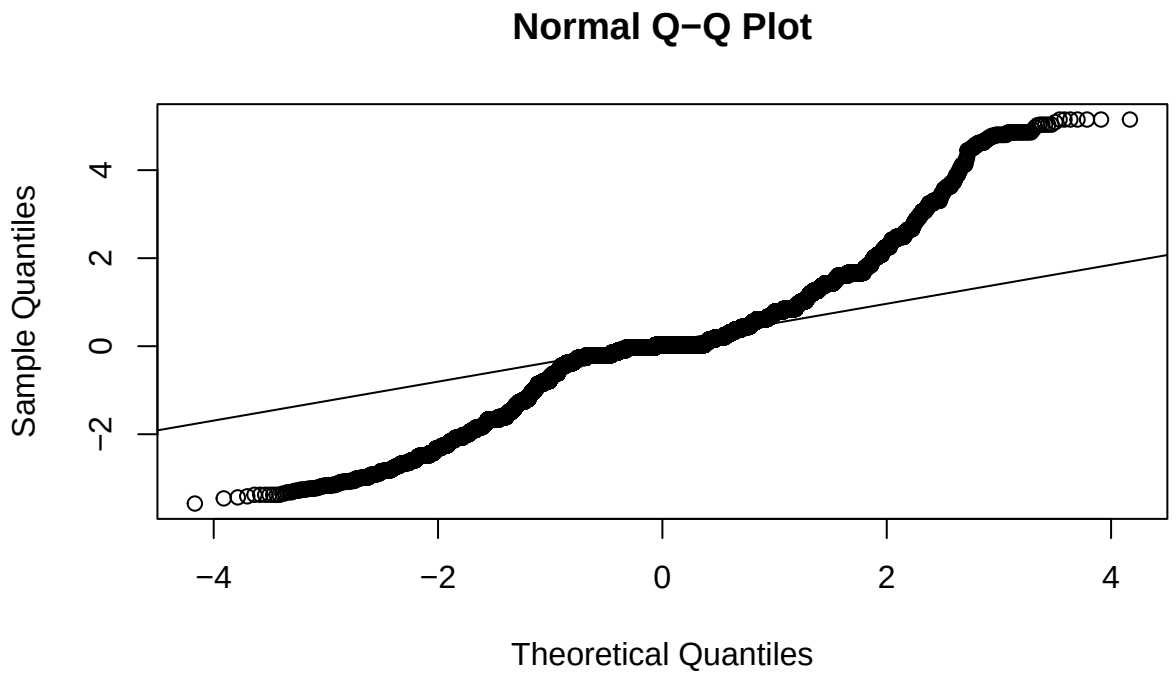
$t_s = 4.7915e + 12$

$p < 2.2e - 16$

Because p $< 0.05$, we can reject the null hypothesis, and conclude that a linear relationship does exist between education level and hours worked. However, the correlation coefficient is suggestive of a very weak positive

linear correlation. We can now try to form a linear regression model for hours worked ~ education level, and test the strength of our model.

```
##
## Call:
## lm(formula = hours ~ education_level)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.645  -2.669   0.331   4.620  62.885
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.2711     0.2737  121.58   <2e-16 ***
## education_level   0.7109     0.0263   27.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.21 on 32559 degrees of freedom
## Multiple R-squared:  0.02194,    Adjusted R-squared:  0.02191
## F-statistic: 730.4 on 1 and 32559 DF,  p-value: < 2.2e-16
```

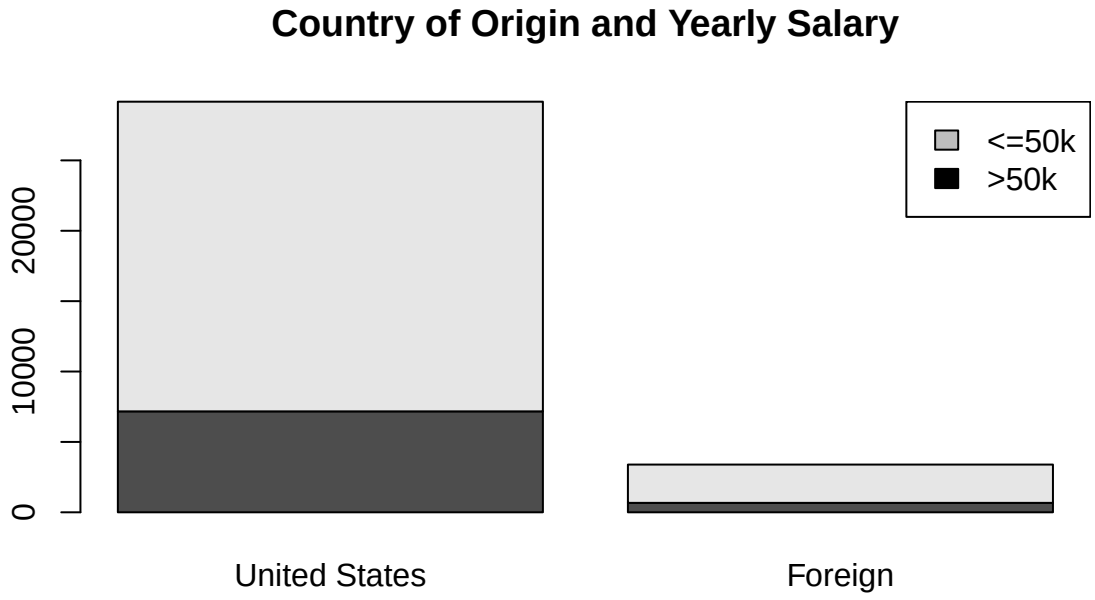



Model Equation: $y = 33.2711 + 0.7109x$

Through this analysis, we can see that the coefficient of determination, R-squared, is equal to 0.022, which represents the proportion of variability in hours worked that is explained by education level. This value is extremely low, which means that our model fit would be extremely inaccurate when predicting hours worked based on education level. This is confirmed by looking at the Normal Q-Q Plot, where a vast majority of the data points do not follow the line y=x.

**Inquiry 6:** Are being from the United States and making more than 50k a year associated?

The variables we will use for this test are "fiftyk", which tells us whether an individual has an annual income higher than 50k, and "country" which will allow us to know whether or not an individual's native country is the United States. When creating a bar plot for these two variables, it seems as though individuals from the United States have a higher success rate of making more than 50k, so our $H_1$ will be that being from the United States and making more than 50k is associated. According to our data set, the amount of individuals who make more than 50k and are from the US is 7,171, the amount of individuals who make less than 50k and are from the US is 21,999, the amount of individuals who make more than 50k and are not from the US is 670, and the amount of individuals who make less than 50k and are not from the US is 2,721. We can use these numbers to create an Expected Contingency Table where we can then utilize a $\chi^2$ Test for Independence.




```
##       [,1] [,2]
## [1,]  7171  670
## [2,] 21999 2721
```

```
##
##  Pearson's Chi-squared test
##
## data:  Income_County
## X-squared = 38.689, df = 1, p-value = 4.97e-10
```

$H_0$ : Being from the United States and making more than 50k a year is independent.

$H_1$ : Being from the United States and making more than 50k is associated.

$\alpha = 0.05$

$\chi^2 = 38.689$

$p = 4.97e - 10$

Using a $\chi^2$ test for independence, we achieve a p-value $< 0.05$, which means we can reject the null hypothesis, and it means we have sufficient evidence to conclude that being from the United States and making more

than 50k is associated.

## Discussion:

While these results alone are not enough to definitively declare whether or not workplace discrimination exists for a certain group of people, it's important to look at the results of our inquiries and think about what may be causing certain trends.

In inquiry 1, we find that men do not have significantly higher levels of education, which means my hypothesis was incorrect. The reasoning behind my hypothesis that men would have a higher level of education is that possibly women would feel societal pressure to not pursue higher education and instead start a family, or they would be discriminated against in the college admissions process all together. I feel that this hypothesis might have held true for data collected in the mid 1900's, but this data represents figures from 1994, around twenty years after Title IX was passed which prohibited sex based discrimination in any educational program that received federal money.

For inquiry 2, my hypothesis that the proportion of non-white workers who make more than 50k is less than the proportion of the general population who makes more than 50k was proven correct. This was not surprising, as it has always been an issue that people of color are not chosen as often for executive or leadership positions that pay much higher wages. Inquiry 4 ties in with inquiry 2, as we found out that the only black workers worked a significantly less amount of hours than white workers, and even then the difference in mean hours worked between those two groups was about 2 hours. With that being said, its a sign of possible discrimination that almost all racial groups work the same amount of weekly hours, yet the proportion of non-white workers who make more than 50k is significantly less than the proportion of the general population that makes 50k.

Inquiry 5 looked at whether there was a linear correlation between education level and hours worked. While we determined that there was in fact a correlation, the correlation was extremely weak and therefore led to an extremely inaccurate linear model. In hindsight, this question is not the most productive question to ask, as almost every American works a 40 hour work week no matter the industry or position. It would have been better to try and predict job-position given education level, but this also might have proven ineffective due to limitations in our data set.

Finally, inquiry 6 tested my hypothesis that being from the United States and making more than 50k is associated. My hypothesis was proven right, which indicates that there might be forms of workplace discrimination towards individuals not from the US, such as not hiring people because they do not speak English as their first language. It would be interesting to compare this data from 1994 to data from today, as the United States has become more and more diverse and has passed more regulation on discriminatory hiring practices.

## Conclusion:

While this research cannot be used to determine what causes discriminatory trends, the results presented in this report help shed light on the various types of discrimination within the workforce, and helps identify what groups of people are at a greater disadvantage when it comes to achieving a high level of education or jobs with higher income. In general, this data analysis shows minorities and people who aren't from the United States are less likely to make as much as their white coworkers despite working close to the same amount of hours. Also, the data indicates that women and men now typically have the same opportunities to advance in education, but women may have a harder time achieving leadership positions, as the age of women in the workforce is on average lower which may be indicative of feeling societal pressure to leave the workforce early. More research should be done on the subject, especially since this data set is representative of the American population in 1994. I would predict that today our workforce would be more diverse than ever, with hopefully more minorities in high-paying leadership roles.

## References:

Data Set from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/census+income

## Appendices:

Code, data set, .names file can be found at https://drive.google.com/drive/folders/1POExcks5Tj0nLE8u-op4ShpvRtmbXUUS?usp=sharing

Video presentation can be found at https://youtu.be/7CmnMr3E618