# EDA on 1994 US Census: An Investigation on Racial Bias

Ian Liu

April 2nd, 2022

## Contents



## Introduction

For centuries, countries have dealt with slavery, segregation, apartheid, and other forms of racial discrimination. Although years have passed after the abolishment of such inhumanities, people and media report racial discrimination in workplaces and other public settings. The term that has been coined is "institutional racism" or "systemic racism." According to Jo Persad, a student at Harvard Graduate School of Education, systemic racism is "this ever-present force, kind of like gravity. You can't see it, but you can experience its effects and it 'holds' the world together." (TODAY.com) Most of those who talk about this issue claim that Caucasians have been privileged in many ways compared to African Americans and other minorities. As a result, many governmental and company policies came forth. Some companies were required to hire a certain amount of Latinos or Hispanics. Almost every job application nowadays requires one's racial and demographic information. Policies like Affirmative Action, a policy to include underrepresented groups, have become a heated topic of debate among politicians and media. The question is, if there is racial discrimination at the institutional level, how much effect does it have?

Through this project we aim to discover how large of an influence race has had on income, education, and work hours of United States individuals. Throughout the entire project, we use the 1994 US Census data set from the UC Irvine Machine Learning Repository. In addition, we use the R packages TidyVerse, ggcorrplot, and gmodels for our data analysis. We start by data cleaning and wrangling, and then advance to the analysis stage. At the conclusion, we offer our understandings and our limitations of the research, so that the reader may continue pursuing the topic on his or her own.

# 1 Data Cleaning & Wrangling

The Census dataset, although informative, has a few inconsistencies with R. First, the missing values are denoted with a "?". In addition, all the values have a space before them. The column names are also nowhere to be found. Let's replace the missing values with NA for convenience, strip away the trailing and leading whitespaces, and give the data set its appropriate column names.

```
## Rows: 32,561
## Columns: 15
## $ Age             <int> 39, 50, 38, 53, 28, 37, 49, 52, 31, 42,~
## $ Workclass       <fct> State-gov, Self-emp-not-inc, Private, P~
## $ fnlwgt          <int> 77516, 83311, 215646, 234721, 338409, 2~
## $ Education       <fct> Bachelors, Bachelors, HS-grad, 11th, Ba~
## $ EducationNumber <int> 13, 13, 9, 7, 13, 14, 5, 9, 14, 13, 10,~
## $ MaritalStatus   <fct> Never-married, Married-civ-spouse, Divo~
## $ Occupation      <fct> Adm-clerical, Exec-managerial, Handlers~
## $ Relationship    <fct> Not-in-family, Husband, Not-in-family, ~
## $ Race            <fct> White, White, White, Black, Black, Whit~
## $ Gender          <fct> Male, Male, Male, Male, Female, Female,~
## $ CapitalGain     <int> 2174, 0, 0, 0, 0, 0, 0, 0, 14084, 5178,~
## $ CapitalLoss     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Hours_per_week  <int> 40, 13, 40, 40, 40, 40, 16, 45, 50, 40,~
## $ NativeCountry   <fct> United-States, United-States, United-St~
## $ Salary          <fct> <=50K, <=50K, <=50K, <=50K, <=50K, <=50~
```

Table ?:

The dataset has 32,561 rows, and 15 columns. Of these 15 columns, 6 are discrete numerical, and the rest are categorical values. For the purpose of this project, we will not require the use of fnlwgt, CapitalGain, CapitalLoss, or NativeCountry.

Now, let's have a look at how many missing values there are, and consider various methodologies to deal with them.

```
##       Age                    Workclass            Education
##  Min.   :17.00   Private         :22696   HS-grad     :10501
##  1st Qu.:28.00   Self-emp-not-inc: 2541   Some-college: 7291
##  Median :37.00   Local-gov       : 2093   Bachelors   : 5355
##  Mean   :38.58   State-gov       : 1298   Masters     : 1723
##  3rd Qu.:48.00   Self-emp-inc    : 1116   Assoc-voc   : 1382
##  Max.   :90.00   (Other)         :  981   11th        : 1175
##                  NA's            : 1836   (Other)     : 5134
##  EducationNumber               MaritalStatus
##  Min.   : 1.00   Divorced              : 4443
##  1st Qu.: 9.00   Married-AF-spouse     :   23
##  Median :10.00   Married-civ-spouse    :14976
##  Mean   :10.08   Married-spouse-absent:  418
##  3rd Qu.:12.00   Never-married         :10683
##  Max.   :16.00   Separated             : 1025
##                  Widowed               :  993
##            Occupation            Relationship
##  Prof-specialty : 4140   Husband       :13193
##  Craft-repair   : 4099   Not-in-family : 8305
##  Exec-managerial: 4066   Other-relative:  981
```

```
##  Adm-clerical   : 3770   Own-child     : 5068
##  Sales          : 3650   Unmarried     : 3446
##  (Other)        :10993   Wife          : 1568
##  NA's           : 1843
##                  Race          Gender      Hours_per_week
##  Amer-Indian-Eskimo:  311   Female:10771   Min.   : 1.00
##  Asian-Pac-Islander: 1039   Male  :21790   1st Qu.:40.00
##  Black             : 3124                  Median :40.00
##  Other             :  271                  Mean   :40.44
##  White             :27816                  3rd Qu.:45.00
##                                            Max.   :99.00
##
##    Salary
##  <=50K:24720
##  >50K : 7841
##
##
##
##
##
```

Tabel ?: A summary of our dataset

Within, the dataset, the variables with the most missing values is Workclass and Occupation. Since they both have a significant amount of NA values, we'll remove them from our data.

```
##   Age        Workclass Education EducationNumber
## 1  39        State-gov Bachelors             13
## 2  50 Self-emp-not-inc Bachelors             13
## 3  38          Private   HS-grad              9
## 4  53          Private      11th              7
## 5  28          Private Bachelors             13
## 6  37          Private   Masters             14
##        MaritalStatus        Occupation  Relationship  Race
## 1      Never-married      Adm-clerical Not-in-family White
## 2 Married-civ-spouse   Exec-managerial       Husband White
## 3           Divorced Handlers-cleaners Not-in-family White
## 4 Married-civ-spouse Handlers-cleaners       Husband Black
## 5 Married-civ-spouse    Prof-specialty          Wife Black
## 6 Married-civ-spouse   Exec-managerial          Wife White
##   Gender Hours_per_week Salary
## 1   Male             40  <=50K
## 2   Male             13  <=50K
## 3   Male             40  <=50K
## 4   Male             40  <=50K
## 5 Female             40  <=50K
## 6 Female             40  <=50K
```

Table ?: Cleaned data

We now end up with 11 columns (3 discrete numerical, 8 categorical), and 30,718 rows that are ready for further exploratory data analysis.

## 2 Racial Bias

Systemic inequality and institutional racism are terms that not only come up during political debates, but also during job and school applications. Those who support this theory often claim that institutional discrimination results from an individual's race or ethnicity. Although segregation supposedly ended in the 1950-1960s, to this day we execute policies that are founded upon the idea that certain races are more advantageous in certain regards.

We will dive into this topic in 3 different ways. First we explore if certain races receive higher education. Next, we look into the correlation of race and income. Lastly, we compare each race's work hours per week.

Let's first look at the distribution across races in our given dataset.

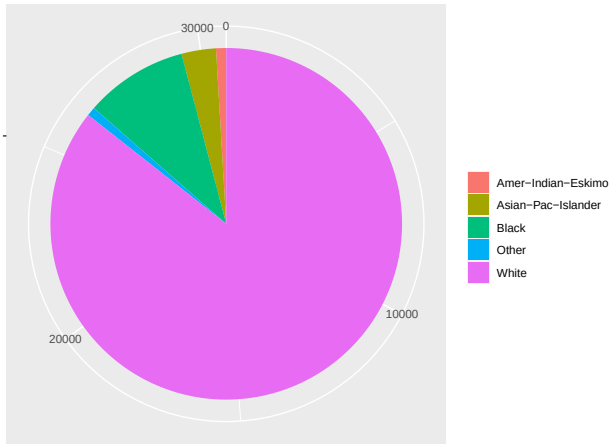


Figure ?: A pie chart of races within our dataset

From Figure ? we can tell that our dataset consists mostly of Whites/Caucasians, followed by: Blacks/African Americans, Asian-Plac-Islander, Amer-Indian-Eskimo, and Other races. Since there's not much we can get from analyzing the "Other" race, we will simply remove it from our observed dataset.

Before we dive into any of the specific categorical variables, let's look at them holistically. The variables we want to visualize are: education, salary, race, and work hours per week. With so many variables involved, it may be best to use a scatterplot to represent the relationships amongst each other. To use a scatterplot to its fullest potential, we need 2 numerical variables as its x and y.

Notice how there is a variable in our table called "EducationNumber," which represents the level of education an individual has attained (Figure ?). We'll use this variable to replace Education.

```
## # A tibble: 16 x 2
## # Groups:   Education [16]
##   Education EducationNumber
##   <fct>               <int>
## 1 Preschool               1
## 2 1st-4th                 2
## 3 5th-6th                 3
## 4 7th-8th                 4
## 5 9th                     5
## 6 10th                    6
## # ... with 10 more rows
```

Figure ?: Education is ranked from preschool to Doctorate (1-16)

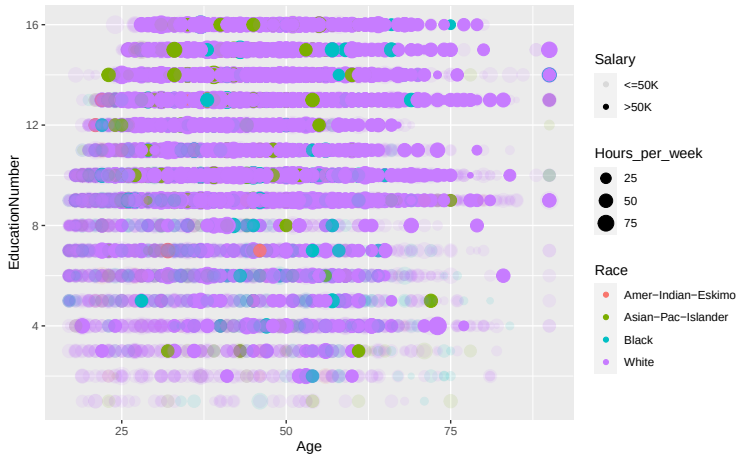


Figure ?: A scatterplot of EducationNumber vs Age, with Salary, Race, and work hours per week represented via transparency, color, and point size.

Immediately we can see that this plot gives us no information whatsoever. All other races have been overshadowed by Caucasian data points, and the hours per week are not clear. To avoid distraction from over representation in our data, we will partition the graph by race into 5 different graphs. Next, we will cube the hours per week so a clear distinction of sizes can be made (Figure ?).

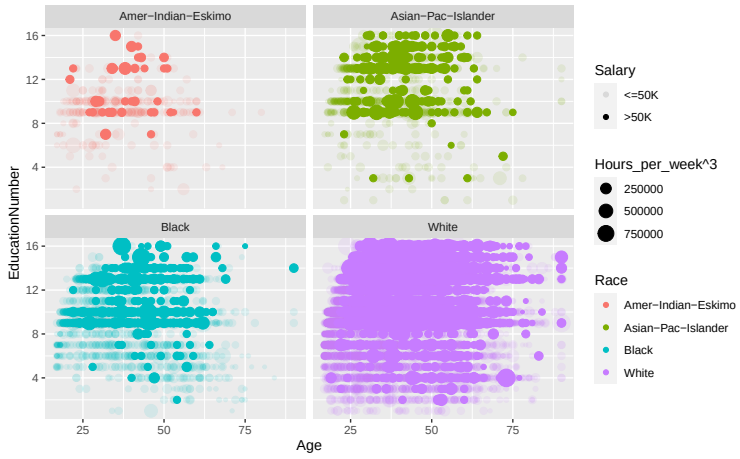


Figure ?: A partition of Figure ?-1's scatterplot by race and exaggerated work hours per week.

Now that we have a pretty good overview of what the data presents, we can delve deeper into each area.

## 2.1 Race & Education

In order create an accurate representation, we can use proportions.

Although we already have a general overview of the connection between race and other variables, there is still an issue. Namely, the data points are not proportions, so we may make the false observation that Caucasians have a higher chance of receiving better education. But this may very well be the result of Caucasians having more representation in the data. In other words, because there are more Blacks and Whites, we will therefore discover more Blacks and Whites among those who have a higher education.

To avoid such distortion, we will create a new table with Race, Salary, and the percentage of people who make that salary within their race (Figure ?).

```
## # A tibble: 6 x 3
## # Groups:   Race [4]
##   Race                EducationNumber percN
##   <fct>                         <int> <dbl>
## 1 Asian-Pac-Islander               16 2.87
## 2 White                            16 1.35
## 3 Amer-Indian-Eskimo               16 0.699
## 4 Black                            16 0.378
## 5 Asian-Pac-Islander               15 4.00
## 6 White                            15 1.89
```

Table ?: A representation of the percentage of people within their race that have the respective education levels.

From Table ? we might infer that Asians actually have a higher chance of being more educated than other races, followed by Whites. Let's visualize this trend.

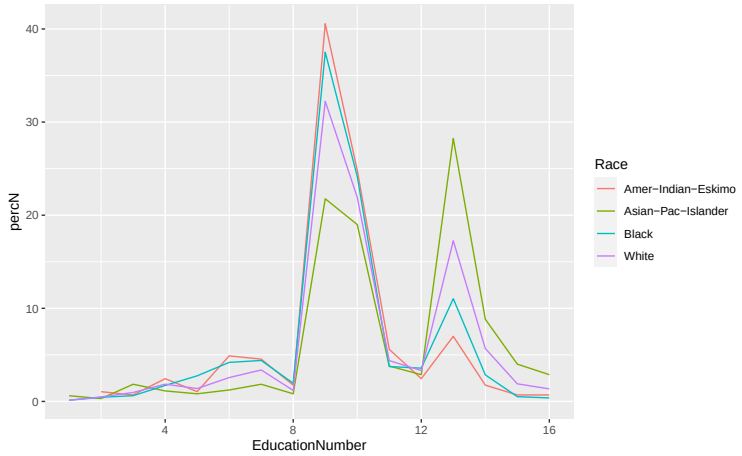


Figure ?: A line plot of education levels vs percentage of those with the specified education level.

As shown in Figure ?, the majority of all races are educated between levels 8-11 (12th to assoc-voc) and a portion is educated in levels 12-14 (Assoc-acdm to Masters). The degree of education most receive is 9, high school degree. Interestingly, Asians see a decline in those who are educated between levels 8-11 (12th to assoc-voc), but see a huge peak in those who are educated in levels 12-14 (Assoc-acdm to Masters). It is also evident that the education gap of other races (besides Asians) is most prominent in levels 12-14 (Assoc-acdm to Masters).

To see how concentrated the values of education levels are, we can use a violin plot overlayed with a box plot (Figure ?).

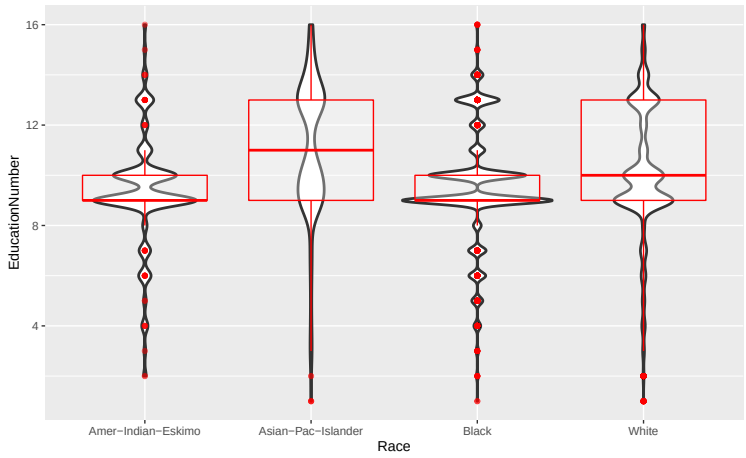


FIGURE ?: A VIOLIN AND BOX PLOT OF EDUCATION NUMBER AND RACES.

Figure? basically confirms with what we have observed so far. Asians and Whites receive better education, while the others fall behind. However, we can also see that Blacks and Amer-Indian-Eskimo groups have smaller boxes, indicating more concentration in the data. Asians and Whites have larger boxes, showing less concentration. Furthermore, Whites have more concentration in the lower half of the box, which means that more people pursue an education level between 9-10 while less pursue an education level of 10-13.

Before we jump to conclusions, we should also bear in mind the effects of other variables. For instance, it is possible that our data contains older or younger people for certain races. The age of a person can possibly have correlation with his or her education, as those who usually obtain higher degrees are slightly older (from personal experience). We can use a correlation matrix to see if they have correlation (Figure ?). Since correlation matricies only work with numerical data, we will only use numerical columns (Age, EducationNumber, Hours_per_week). For future convenience, we've also added a new column SalNum to our data set, which is 50,000 if the Salary column is "<=50K", and 51,000 if the Salary column is ">50K."

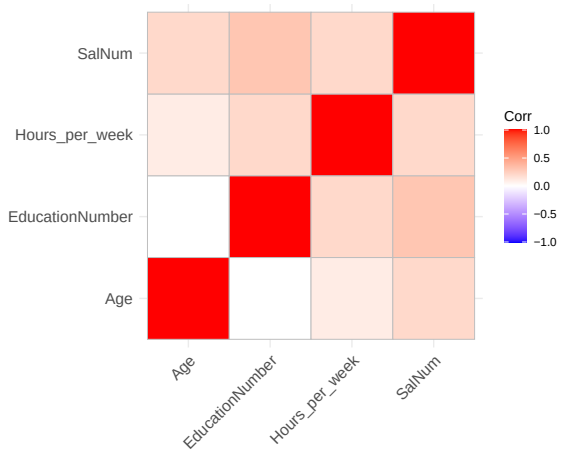


FIGURE ?: A CORRELATION MATRIX OF AGE, EDUCATION, WORK HOURS PER WEEK, AND SALARY. THE DARKER THE COLOR, THE MORE THE CORRELATION. RED = POSITIVE CORRELATION, AND BLUE = NEGATIVE CORRELATION

Oddly enough, we see 0 correlation of age and education. Since the majority of our data are high school graduates (Figure lineplot), let's try limiting our data only to those who've received a high school diploma or higher degree. Even after we filter our data, the updated correlation matrix (Figure 2nd corr matrix), shows small correlation between age and education. In this case, such small correlation can probably be ignored.

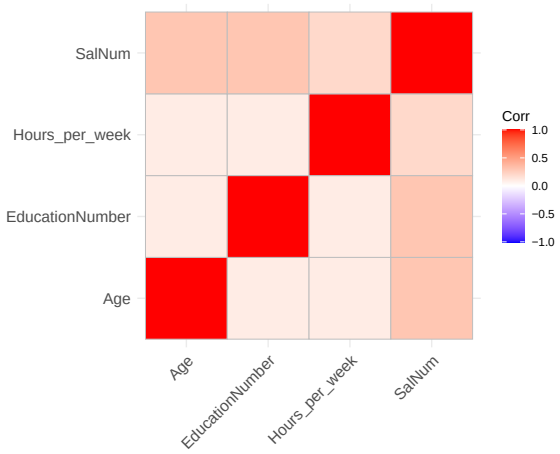


Figure ?: A correlation matrix of age, education, work hours per week, and salary for people with an education level of 9 or above (high school grad or higher). The darker the color, the more the correlation. Red = positive correlation, and blue = negative correlation

To double check, let's examine the distribution of ages across races (Figure ). The distributions are roughly similar, and there is likely no reason to believe that the education level difference across races is due to the difference in age distributions.

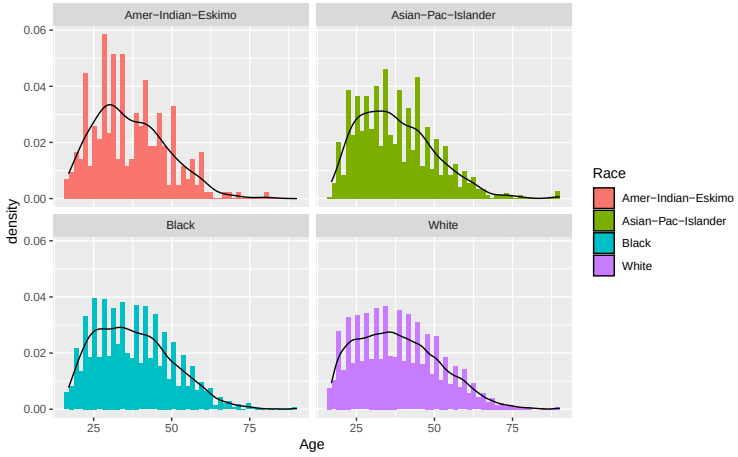


Figure ?: Histograms overlayed with density plots of education levels across ages for different races.

## 2.2 Race & Income

To analyze race and income, we should start off by, again, looking at the raw data, then use proportions to accurately determine if race and income may have any correlation.

```
##
##
##    Cell Contents
## |-------------------------|
```

```
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  30470
##
##
##                    | raceAdultData$Salary
## raceAdultData$Race |     <=50K |      >50K | Row Total |
## -------------------|-----------|-----------|-----------|
## Amer-Indian-Eskimo |       252 |        34 |       286 |
##                    |     6.590 |    19.737 |           |
##                    |     0.881 |     0.119 |     0.009 |
##                    |     0.011 |     0.004 |           |
##                    |     0.008 |     0.001 |           |
## -------------------|-----------|-----------|-----------|
## Asian-Pac-Islander |       703 |       271 |       974 |
##                    |     1.013 |     3.034 |           |
##                    |     0.722 |     0.278 |     0.032 |
##                    |     0.031 |     0.036 |           |
##                    |     0.023 |     0.009 |           |
## -------------------|-----------|-----------|-----------|
##              Black |      2531 |       378 |      2909 |
##                    |    56.221 |   168.384 |           |
##                    |     0.870 |     0.130 |     0.095 |
##                    |     0.111 |     0.050 |           |
##                    |     0.083 |     0.012 |           |
## -------------------|-----------|-----------|-----------|
##              White |     19357 |      6944 |     26301 |
##                    |     6.593 |    19.746 |           |
##                    |     0.736 |     0.264 |     0.863 |
##                    |     0.847 |     0.910 |           |
##                    |     0.635 |     0.228 |           |
## -------------------|-----------|-----------|-----------|
##       Column Total |     22843 |      7627 |     30470 |
##                    |     0.750 |     0.250 |           |
## -------------------|-----------|-----------|-----------|
##
##
```

Table ?: A contingency table

```
## # A tibble: 8 x 3
## # Groups:   Race [4]
##   Race               Salary percN
##   <fct>              <fct>  <dbl>
## 1 Asian-Pac-Islander >50K    27.8
## 2 White              >50K    26.4
## 3 Black              >50K    13.0
## 4 Amer-Indian-Eskimo >50K    11.9
```

```
## 5 Amer-Indian-Eskimo <=50K   88.1
## 6 Black              <=50K   87.0
## # ... with 2 more rows
```

Figure ?

Table ? shows that our data consists mostly of those who earn less than 50 thousand and an large portion of Whites. From Figure? it is already obvious that there is a huge income gap between Blacks and Whites.

Even though our current observations suggest that certain races make more, it does not prove that they make more because of discrimination. So far, the races that make the least are also the ones that are the least educated. Since some races receive poorer education, the fact that they make less could very likely be purely caused by their lack of sufficient education.

To take into account all of the possible reasons why one race might earn more than the other, we can look at each variable's effect on income separately.

### 2.2.1 Observing Income Difference Across Races for the Same Education Levels

Education levels has a strong correlation with income, as shown in Fig?. Here, we start by creating 2 new tables. Table ? displays how many people within a race receive a certain education level. Table ? details the number of people who receive a certain salary in a given education level and race. By combining information from the 2 tables, we form Table ?, which contains Table ?'s new column divide by Table ?'s new column, giving us the percentage of people in that education level and in that race who receive the given salary.

```
## # A tibble: 6 x 3
## # Groups:   Race [1]
##   Race               Education raceedtotal
##   <fct>              <fct>           <int>
## 1 Amer-Indian-Eskimo 10th               14
## 2 Amer-Indian-Eskimo 11th               13
## 3 Amer-Indian-Eskimo 12th                5
## 4 Amer-Indian-Eskimo 1st-4th             3
## 5 Amer-Indian-Eskimo 5th-6th             2
## 6 Amer-Indian-Eskimo 7th-8th             7
```

Table ?: The number of people in each race as shown in racetotal

```
## # A tibble: 6 x 4
## # Groups:   Race, Education [5]
##   Race               Education Salary     n
##   <fct>              <fct>     <fct>  <int>
## 1 Amer-Indian-Eskimo 10th      <=50K     14
## 2 Amer-Indian-Eskimo 11th      <=50K     11
## 3 Amer-Indian-Eskimo 11th      >50K       2
## 4 Amer-Indian-Eskimo 12th      <=50K      5
## 5 Amer-Indian-Eskimo 1st-4th   <=50K      3
## 6 Amer-Indian-Eskimo 5th-6th   <=50K      2
```

Table ?: The number of people in each race with the specified education and salary

```
##                 Race Education raceedtotal Salary   n     propor
## 1 Amer-Indian-Eskimo Bachelors          20  <=50K  12 0.60000000
```

```
## 2 Amer-Indian-Eskimo Bachelors          20   >50K   8 0.40000000
## 3 Amer-Indian-Eskimo Doctorate           2  <=50K   1 0.50000000
## 4 Amer-Indian-Eskimo Doctorate           2   >50K   1 0.50000000
## 5 Amer-Indian-Eskimo   HS-grad         116  <=50K 105 0.90517241
## 6 Amer-Indian-Eskimo   HS-grad         116   >50K  11 0.09482759
```

Table ?: A merged table from Table? and Table? with propor representing the proportion of people in a race that have a level of education and salary

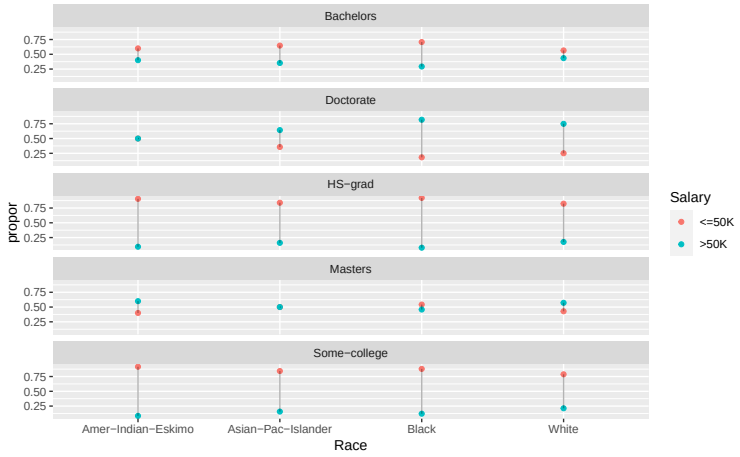


Figure ?: A dot plot of Table ?'s propor column. The taller the line, the larger the difference in proportion of those who make >50K and those who make <=50K.

Figure ? details the salary level of races in certain education levels.

- **High school grads**: We see that most of them earn less than 50k. Asians and Whites have noticeably less percentage of people who make less than 50k, and more who earn more than 50k. Although this difference is small, it shows that within a group of high school graduates, Asians and Whites are slightly more likely to make more money. (However this could be because other races are more likely to choose lower-paying occupations)
- **Some college**: There is not much of difference salary-wise among races except for Whites. Whites are more likely to be paid higher in this education category.
- **Bachelors**: About more than half the people with a Bachelor's degree earn <50k, but nearly half also earn >50k. White and Amer-Indian-Eskimos have a smaller proportion difference between those who earn <50k and those who earn >50k. Asians and especially Blacks, have a much larger difference in this regard. Again, this may be the result of preferred occupations for their respective races.
- **Masters**: Individuals with a Master's degree end up having just about equal proportions for those who earn more and those who earn less. There is not much variation for any races.
- **Doctorate**: At the Doctorate level, the proportion of those who make more than 50k are higher than those who make less than 50k. Contrary to popular belief, African Americans with a Doctorate degree actually have an upper hand in their job search. Following after Blacks is Whites, who also have noticeable more individuals that earn >50K compared to Asian-Pacific-Islanders and American-Indian-Eskimos.

From the above, we can see that after comparing those within the same education level, the compensation is roughly similar. Only at the Doctorate level do things work a bit more in favor towards Blacks and Whites. Since the difference is not humongously large, it could also be attributed to Whites and Blacks having more representation in our data.

### 2.2.2 Observing Income Difference Across Races for the Same Ages

As we saw earlier in the correlation matrix (Fig ?), age and income have a pretty strong correlation, perhaps just as much as does education. We will map out the density curves of both income levels for each race (Fig?). A general trend we see is that those who earn more, tend to be older (Fig ?). The age distributions of those who make more and those who make less across races are also approximately equal. We can probably say that: although age makes a difference on income, the age distribution is relatively similar regardless of race, so the effect of age on income can be neglected when comparing income across races.

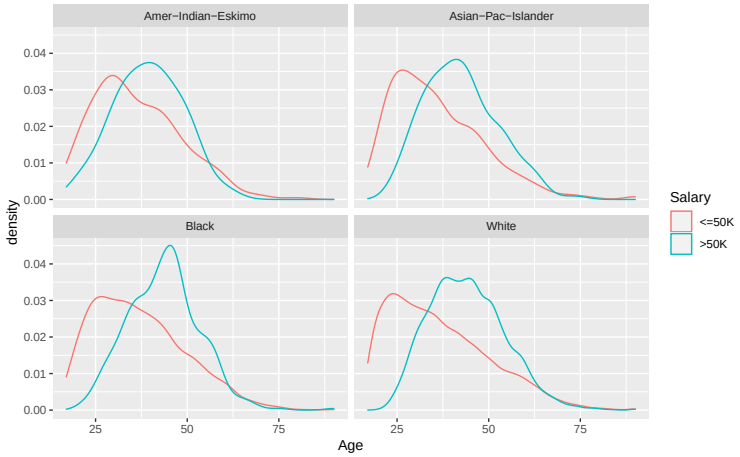


Figure ?: A density plot of age and proportion of people in that race with the age. Black represents the density curve of all individuals, while blue and red represent the density plot for each income level.

### 2.2.3 Observing Income Difference Across Races for the Same Occupations

Lastly, we may observe income differences across races for the same occupations. From the results of Figure? we can see that the number of occupations is so large that it is difficult to even wrap our heads. We may need more data to further investigate if a certain race has a favor in certain occupation(s). And if so, how much, if at all, does it affect income.

```
##                  Race        Occupation Salary raceOccuN raceN
## 1  Amer-Indian-Eskimo      Adm-clerical  <=50K        28   286
## 2  Amer-Indian-Eskimo      Adm-clerical   >50K         3   286
## 3  Amer-Indian-Eskimo      Armed-Forces  <=50K         1   286
## 4  Amer-Indian-Eskimo      Craft-repair  <=50K        38   286
## 5  Amer-Indian-Eskimo      Craft-repair   >50K         6   286
## 6  Amer-Indian-Eskimo   Exec-managerial  <=50K        27   286
## 7  Amer-Indian-Eskimo   Exec-managerial   >50K         3   286
## 8  Amer-Indian-Eskimo   Farming-fishing  <=50K        10   286
## 9  Amer-Indian-Eskimo Handlers-cleaners  <=50K        22   286
## 10 Amer-Indian-Eskimo Machine-op-inspct  <=50K        19   286
## 11 Amer-Indian-Eskimo     Other-service  <=50K        31   286
## 12 Amer-Indian-Eskimo     Other-service   >50K         2   286
## 13 Amer-Indian-Eskimo    Prof-specialty  <=50K        22   286
## 14 Amer-Indian-Eskimo    Prof-specialty   >50K        11   286
## 15 Amer-Indian-Eskimo   Protective-serv  <=50K         6   286
## 16 Amer-Indian-Eskimo   Protective-serv   >50K         2   286
```

```
## 17 Amer-Indian-Eskimo             Sales  <=50K        22   286
## 18 Amer-Indian-Eskimo             Sales   >50K         4   286
## 19 Amer-Indian-Eskimo      Tech-support  <=50K         4   286
## 20 Amer-Indian-Eskimo  Transport-moving  <=50K        22   286
## 21 Amer-Indian-Eskimo  Transport-moving   >50K         3   286
## 22 Asian-Pac-Islander      Adm-clerical  <=50K       117   974
## 23 Asian-Pac-Islander      Adm-clerical   >50K        22   974
## 24 Asian-Pac-Islander      Craft-repair  <=50K        64   974
## 25 Asian-Pac-Islander      Craft-repair   >50K        25   974
## 26 Asian-Pac-Islander   Exec-managerial  <=50K        74   974
## 27 Asian-Pac-Islander   Exec-managerial   >50K        61   974
## 28 Asian-Pac-Islander   Farming-fishing  <=50K        14   974
## 29 Asian-Pac-Islander   Farming-fishing   >50K         2   974
## 30 Asian-Pac-Islander Handlers-cleaners  <=50K        22   974
## 31 Asian-Pac-Islander Handlers-cleaners   >50K         1   974
## 32 Asian-Pac-Islander Machine-op-inspct  <=50K        49   974
## 33 Asian-Pac-Islander Machine-op-inspct   >50K        10   974
## 34 Asian-Pac-Islander     Other-service  <=50K       111   974
## 35 Asian-Pac-Islander     Other-service   >50K        17   974
## 36 Asian-Pac-Islander   Priv-house-serv  <=50K         4   974
## 37 Asian-Pac-Islander    Prof-specialty  <=50K        95   974
## 38 Asian-Pac-Islander    Prof-specialty   >50K        91   974
## 39 Asian-Pac-Islander   Protective-serv  <=50K        10   974
## 40 Asian-Pac-Islander   Protective-serv   >50K         5   974
## 41 Asian-Pac-Islander             Sales  <=50K        87   974
## 42 Asian-Pac-Islander             Sales   >50K        21   974
## 43 Asian-Pac-Islander      Tech-support  <=50K        32   974
## 44 Asian-Pac-Islander      Tech-support   >50K        12   974
## 45 Asian-Pac-Islander  Transport-moving  <=50K        24   974
## 46 Asian-Pac-Islander  Transport-moving   >50K         4   974
## 47              Black      Adm-clerical  <=50K       448  2909
## 48              Black      Adm-clerical   >50K        42  2909
## 49              Black      Armed-Forces  <=50K         1  2909
## 50              Black      Craft-repair  <=50K       195  2909
## 51              Black      Craft-repair   >50K        49  2909
## 52              Black   Exec-managerial  <=50K       160  2909
## 53              Black   Exec-managerial   >50K        84  2909
## 54              Black   Farming-fishing  <=50K        42  2909
## 55              Black Handlers-cleaners  <=50K       168  2909
## 56              Black Handlers-cleaners   >50K        11  2909
## 57              Black Machine-op-inspct  <=50K       252  2909
## 58              Black Machine-op-inspct   >50K        22  2909
## 59              Black     Other-service  <=50K       554  2909
## 60              Black     Other-service   >50K        17  2909
## 61              Black   Priv-house-serv  <=50K        28  2909
## 62              Black    Prof-specialty  <=50K       173  2909
## 63              Black    Prof-specialty   >50K        66  2909
## 64              Black   Protective-serv  <=50K        77  2909
## 65              Black   Protective-serv   >50K        25  2909
## 66              Black             Sales  <=50K       223  2909
## 67              Black             Sales   >50K        31  2909
## 68              Black      Tech-support  <=50K        58  2909
## 69              Black      Tech-support   >50K        13  2909
## 70              Black  Transport-moving  <=50K       152  2909
```

```
## 71              Black  Transport-moving   >50K        18  2909
## 72              White      Adm-clerical  <=50K      2645 26301
## 73              White      Adm-clerical   >50K       439 26301
## 74              White      Armed-Forces  <=50K         6 26301
## 75              White      Armed-Forces   >50K         1 26301
## 76              White      Craft-repair  <=50K      2850 26301
## 77              White      Craft-repair   >50K       844 26301
## 78              White   Exec-managerial  <=50K      1828 26301
## 79              White   Exec-managerial   >50K      1818 26301
## 80              White   Farming-fishing  <=50K       802 26301
## 81              White   Farming-fishing   >50K       113 26301
## 82              White Handlers-cleaners  <=50K      1060 26301
## 83              White Handlers-cleaners   >50K        74 26301
## 84              White Machine-op-inspct  <=50K      1394 26301
## 85              White Machine-op-inspct   >50K       217 26301
## 86              White     Other-service  <=50K      2422 26301
## 87              White     Other-service   >50K       101 26301
## 88              White   Priv-house-serv  <=50K       113 26301
## 89              White   Priv-house-serv   >50K         1 26301
## 90              White    Prof-specialty  <=50K      1969 26301
## 91              White    Prof-specialty   >50K      1682 26301
## 92              White   Protective-serv  <=50K       341 26301
## 93              White   Protective-serv   >50K       178 26301
## 94              White             Sales  <=50K      2313 26301
## 95              White             Sales   >50K       924 26301
## 96              White      Tech-support  <=50K       548 26301
## 97              White      Tech-support   >50K       258 26301
## 98              White  Transport-moving  <=50K      1066 26301
## 99              White  Transport-moving   >50K       294 26301
##           propR
## 1  9.790210e-02
## 2  1.048951e-02
## 3  3.496503e-03
## 4  1.328671e-01
## 5  2.097902e-02
## 6  9.440559e-02
## 7  1.048951e-02
## 8  3.496503e-02
## 9  7.692308e-02
## 10 6.643357e-02
## 11 1.083916e-01
## 12 6.993007e-03
## 13 7.692308e-02
## 14 3.846154e-02
## 15 2.097902e-02
## 16 6.993007e-03
## 17 7.692308e-02
## 18 1.398601e-02
## 19 1.398601e-02
## 20 7.692308e-02
## 21 1.048951e-02
## 22 1.201232e-01
## 23 2.258727e-02
## 24 6.570842e-02
```

```
## 25 2.566735e-02
## 26 7.597536e-02
## 27 6.262834e-02
## 28 1.437372e-02
## 29 2.053388e-03
## 30 2.258727e-02
## 31 1.026694e-03
## 32 5.030801e-02
## 33 1.026694e-02
## 34 1.139630e-01
## 35 1.745380e-02
## 36 4.106776e-03
## 37 9.753593e-02
## 38 9.342916e-02
## 39 1.026694e-02
## 40 5.133470e-03
## 41 8.932238e-02
## 42 2.156057e-02
## 43 3.285421e-02
## 44 1.232033e-02
## 45 2.464066e-02
## 46 4.106776e-03
## 47 1.540048e-01
## 48 1.443795e-02
## 49 3.437607e-04
## 50 6.703334e-02
## 51 1.684428e-02
## 52 5.500172e-02
## 53 2.887590e-02
## 54 1.443795e-02
## 55 5.775180e-02
## 56 3.781368e-03
## 57 8.662771e-02
## 58 7.562736e-03
## 59 1.904435e-01
## 60 5.843933e-03
## 61 9.625301e-03
## 62 5.947061e-02
## 63 2.268821e-02
## 64 2.646958e-02
## 65 8.594019e-03
## 66 7.665865e-02
## 67 1.065658e-02
## 68 1.993812e-02
## 69 4.468890e-03
## 70 5.225163e-02
## 71 6.187693e-03
## 72 1.005665e-01
## 73 1.669138e-02
## 74 2.281282e-04
## 75 3.802137e-05
## 76 1.083609e-01
## 77 3.209003e-02
## 78 6.950306e-02
```

```
## 79 6.912285e-02
## 80 3.049314e-02
## 81 4.296415e-03
## 82 4.030265e-02
## 83 2.813581e-03
## 84 5.300179e-02
## 85 8.250637e-03
## 86 9.208775e-02
## 87 3.840158e-03
## 88 4.296415e-03
## 89 3.802137e-05
## 90 7.486407e-02
## 91 6.395194e-02
## 92 1.296529e-02
## 93 6.767804e-03
## 94 8.794342e-02
## 95 3.513174e-02
## 96 2.083571e-02
## 97 9.809513e-03
## 98 4.053078e-02
## 99 1.117828e-02
```

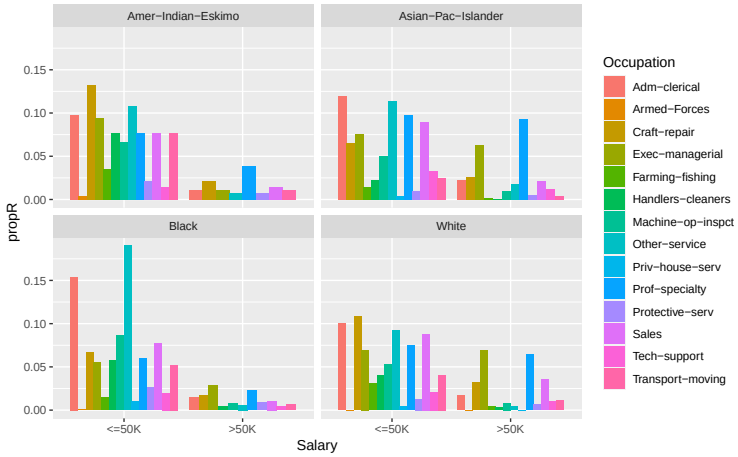


Figure ?

## 2.3 Race & Work hours

To decide if race has a significant role on work hours, we use a histogram overlayed with a density plot (Figure). The plot shows that most of the races work an average 40 hours per week, which is usually the case for a 9-5 job working 5 days a week (8 hours a day multiplied by 5). Although there may not seem to be a discrepancy, it might be slightly strange that the difference in salary across races didn't affect each race's number of work hours.

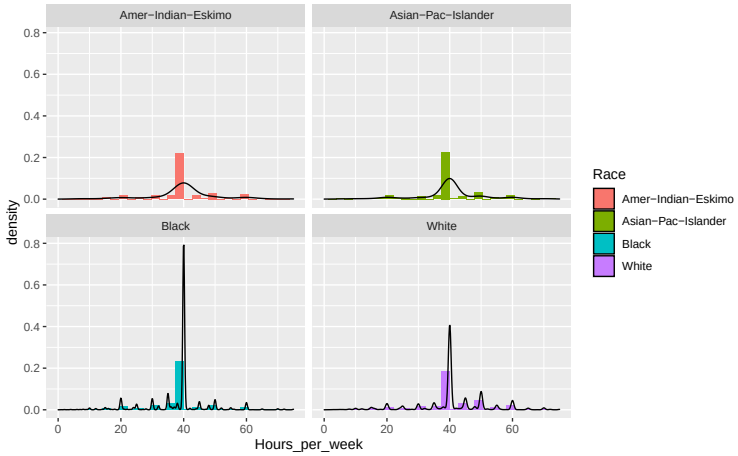


FIGURE ?:

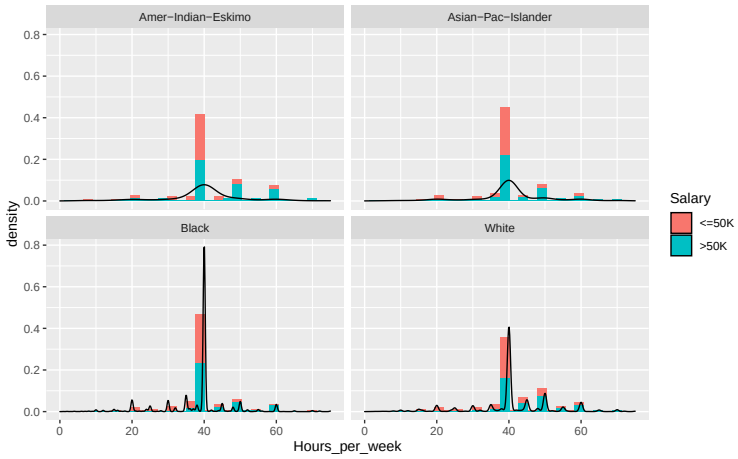


FIGURE ?:

Through Figure ?, we make the following discoveries:

- Those who work more hours tend to be those who make more than 50k
- Despite salary, most people work around 40 hours a week, showing that the correlation between work hours and salary may not be as much as imagined
- The majority of all races have a lower salary, so the difference of the most frequent work hours is minor

One of the most advertised pathways to a financially-stable haven is education. Another claim is that certain races are paid more and are more educated, which is often used as evidence for systemic inequality. We start by exploring the validity of these claims, as well as possible costs that may come with a higher income. Additionally, we attempt to see if age plays an significant role in the determination of salary or education.

# Conclusion

Throughout this exploratory data analysis, we explored areas of racial bias on the education, income, and work-hour levels.

- Work hours: We have found that each race has about identical working hours. In other words, even if there was racial discrimination at workplaces, it was probably not achieved by adding work hours
- Education: From the data, we can conclude that certain races are more educated. We also considered that age distribution might affect each race's education distribution. However, upon further investigation, we noticed that age affected each race in similar ways and did not have as much of an effect on our previous conclusion.
- Income: We noticed that certain races like Asian-Pacific-Islanders and Whites tend to earn more. However, correlation doesn't mean causation. We delved into 3 areas that may possibly influence income:
  - Occupation: Occupation was more complex since it could be affected by workclass. So if we were to notice a specific race working more in a certain occupation, it could be the result of insufficient education, workclass, or other variables. To reach a concrete conclusion, we may need to factor in a lot more other variables as well.
  - Education: Education showed strong correlation with income in the correlation matrix. Therefore, it could easily change our conclusion. We realized that, after equalizing the education level and comparing different races, each race had about the same proportion of people making more money given the same education level. If anything, African Americans made more at the Doctorate Degree level.
  - Age: After noticing some strong correlation between income and age, we looked further. Again, like it was with our education analysis, age affected races in similar ways since each race had nearly identical age distributions, regardless of their salary level.

From our observations, we've come to the realization that none of our findings can support the theory of institutional/systemic racism. However, this does not mean it doesn't exist. Our analyses of the data was limited in many different ways. For instance, we weren't able to consider the effect of other variables on race. Working class, relationships, and others could all have a certain pattern in a certain race. These trends may indirectly or directly influence our findings. Regardless of this, our research suggests that by looking solely at, income, education, age, and working hours, we are unable to conclude unequivocally that systemic/institutional racial discrimination exists or doesn't exist.

In order to see if institutional racism is at work, we may need more complex data that for example, tells us how much their actual salary is. The problem with having salary as a categorical variable is that we don't know how far apart the salary differences are, and we allow those who make 100k per year to share a category with those that make 51k. These small differences add up to a lot when it's a variable we would like to test.

If we were to dive further into the realm of this research, we would have to use a lot more variables such as native countries, marital status, etc. to further refine our research.