Race, Income, and Education

Afif Mazhar

November 13, 2020

Introduction

Exploring data provided by the Census on race, education level, income, and poverty levels, we seek to understand the relationship they share across dissimilar demographic areas. The analysis and prospective recommendation are included in this report. The Census conducted by the United States is used to display not only demographic information about the general population, but provides key insights when combined with corresponding data about income and education, which is vital to making legislative decisions, thus fulfilling the inherent purpose of creating Census data. By exploring this information between and within the States, bound by expected median income and education levels, establishing relationships amid this dataset is imperative to finding those in need of further support for public education, if any, or if public funding disparities continue to exist among ethnic groups. Various data analysis techniques were applied in order to establish if there is a relationship between the location and status of certain population groups, but more importantly, where certain groups continue to experience fundamental disparities compared to others.

Background

The on-going national strife between political factions has sparked the controversial issue of racial disparity prevailing as a dominant factor in minority progress. Heterogeneous groups with particular philosophies raise the question of whether or not every American truly lies on an equal playing field. Furthermore, US citizens universally claim that they are being treated differently according to the color of their skin. The de facto argument and most commonly used rhetoric

summarizes that there lies a cardinal correlation between race and socioeconomic status. Reginald A. Noël, Economist at the Department of Labor further validates this notion in his 2018 article "Race, Economics, and Social Status" by claiming that race is one of many factors that play a critical role in defining social status (Noel 2018). Socioeconomic status is fundamentally acknowledged in modern society; however, minorities may be several steps behind when it comes down to analyzing their position in the status quo.

In order to explore this claim further, we additionally offer a possible warrant for analyzing the income and the subsequent poverty-level affiliations with race: education. The argument we hypothesize is that race and education play a multicausal role in determining levels of poverty between differing groups in the US. Our motivation for utilizing this dataset is to understand, if there exists, the correlation between education and race to the impact of overall income. This is pertinent because interpreting such disparities can give researchers a more definitive approach to public policy and solving race-based issues.

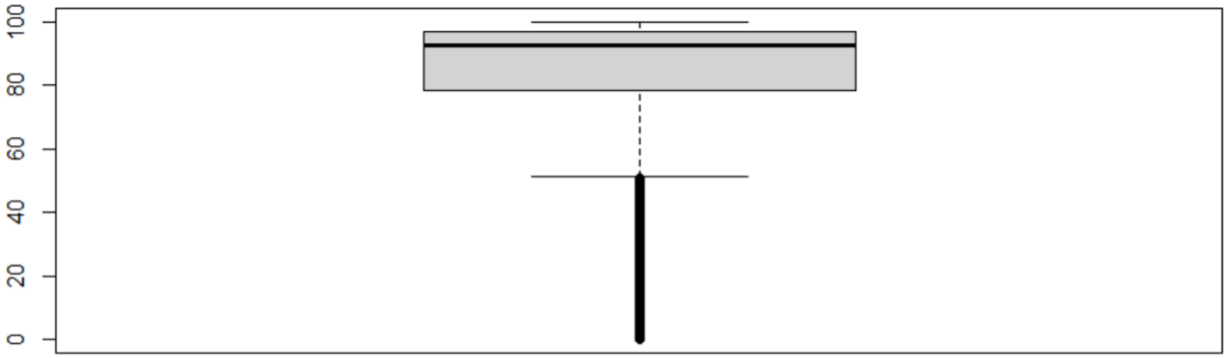
Data

We will be using the second-hand data from 2015 compiled by the Washington Post. The dataset consists of U.S Census information displaying the variables: Geographic area, City, percent White, percent Black, percent Native American, percent Asian, percent Hispanic, along with each city's median household income, the graduation rate of cities with a population over 25 people, and percent below poverty level. We selected this dataset not only to highlight the different race percentages that exist per state and city, but to explain the level of which demographics play a role in poverty and consequently education. The data is consistently valid with minor null values, but overall variables are complete. Figure 1 below displays the summary statistics of each column in the specific dataset. Each column portrays a typical mean which we use as the standard for evaluating the data at hand; i.e., the share_white column has a mean of 83.19% -- meaning that 83.19% of the population in a typical city is White.

```
 share_white      share_black
Min.   :  0.00   Min.   :  0.000
1st Qu.: 78.50   1st Qu.:  0.200
Median : 92.40   Median :  0.800
Mean   : 83.19   Mean   :  7.007
3rd Qu.: 96.80   3rd Qu.:  4.500
Max.   :100.00   Max.   :100.000
share_native_american  share_asian
Min.   :  0.000       Min.   : 0.000
1st Qu.:  0.100       1st Qu.: 0.000
Median :  0.300       Median : 0.400
Mean   :  2.697       Mean   : 1.602
3rd Qu.:  0.800       3rd Qu.: 1.200
Max.   :100.000       Max.   :67.100
share_hispanic    Median.Household.Income
Min.   :  0.000   Min.   :     0
1st Qu.:  1.200   1st Qu.: 35388
Median :  2.900   Median : 45115
Mean   :  8.834   Mean   : 51072
3rd Qu.:  7.900   3rd Qu.: 59310
Max.   :100.000   Max.   :244083
X..over.25.completed.hs X..below.poverty.level
Min.   :  0.00          Min.   :  0.00
1st Qu.: 81.00          1st Qu.:  7.60
Median : 88.30          Median : 14.00
Mean   : 85.86          Mean   : 16.38
3rd Qu.: 93.30          3rd Qu.: 22.60
Max.   :100.00          Max.   :100.00
```

Figure 1

Figures 2 - 5 are excerpts of the dataset where we evaluate individual columns respective to their percentage value. For example, Figure 2 depicts a boxplot of the share_White column where 83.19% of the population in an average city is White. Comparatively, Figure 3 (7.0007%), Figure 4 (2.697%), and Figure 5 (1.602%) are all significantly lower in their typical averages.



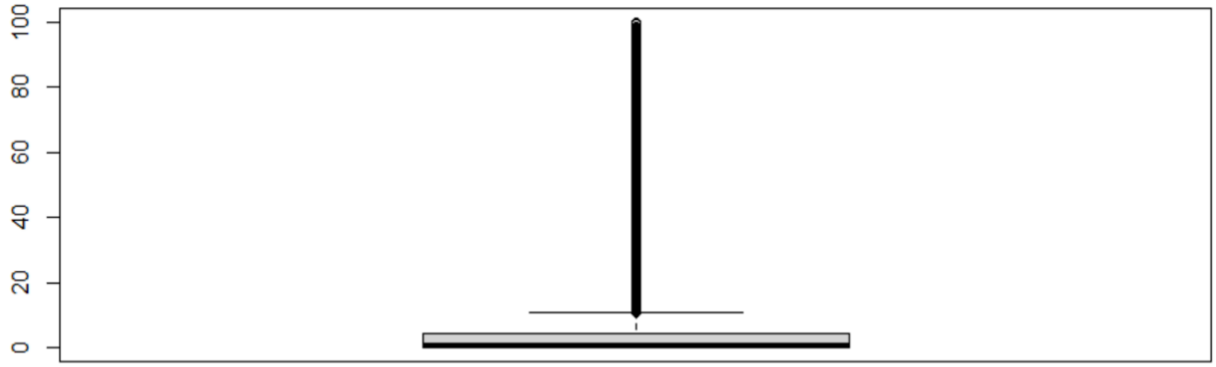


Figure 3

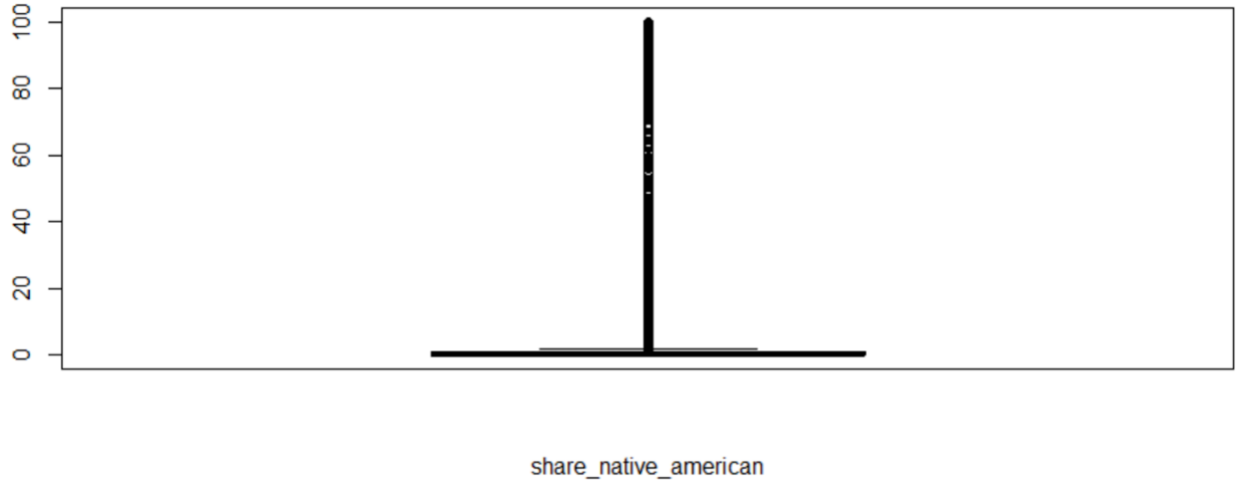


Figure 4

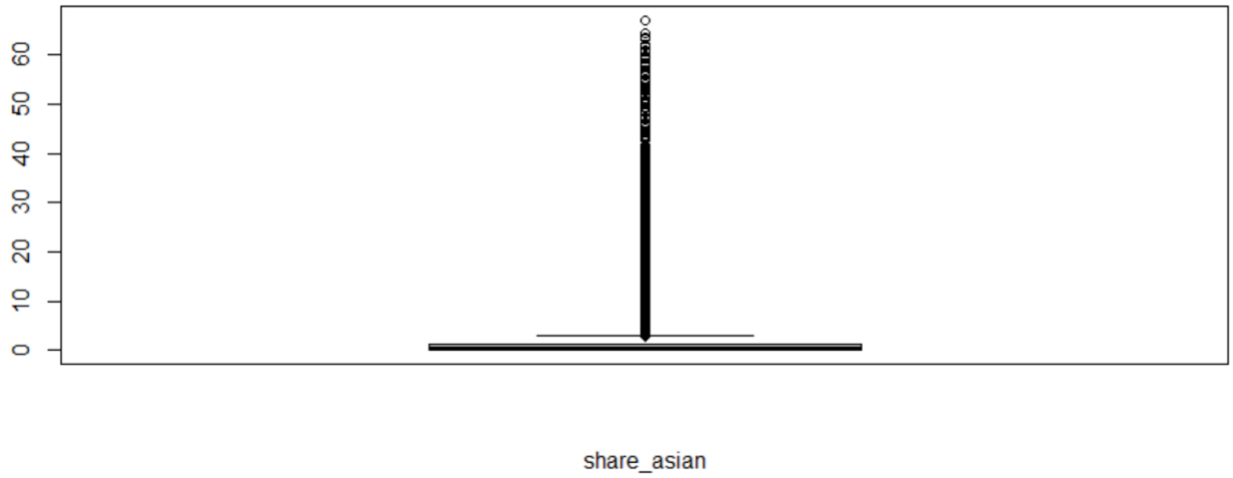


Figure 5

Figure 6 shows a large variation of Median Household Income, with a range between 0 to 250,000. A median household income of 50,000 occurs with the largest frequency within our dataset. The histogram is positively skewed, towards the right.

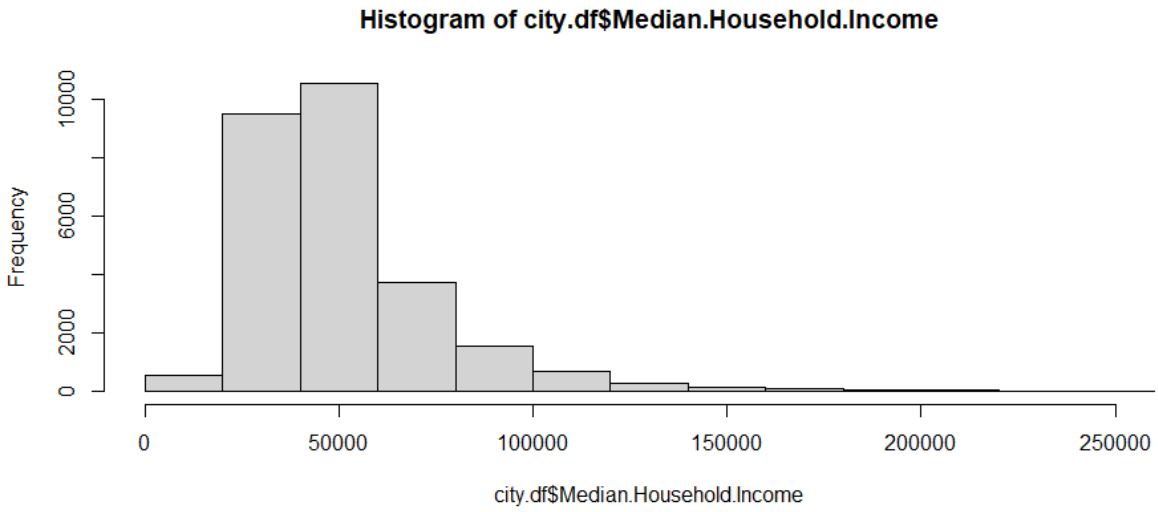


Figure 6

Figure 7 displays the relationship between two variables, Median Household Income (our target variable) and percent over 25 that have completed highschool. It shows a positive relationship between the two, as more of the population completed high school, the higher the median household income within that area.

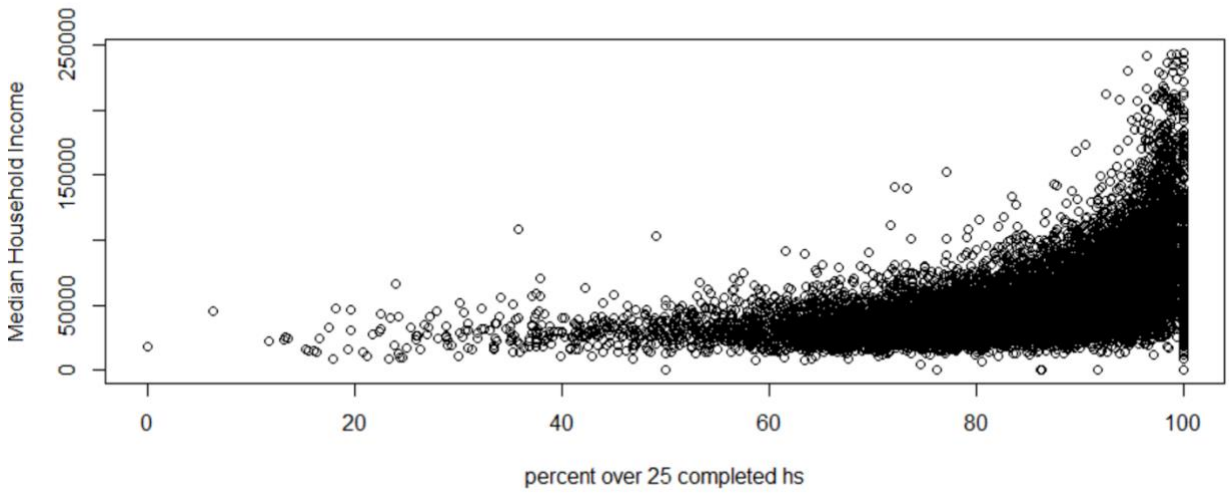


Figure 7

Model Selection

We decided to utilize multiple linear regression for our data analysis model, which is categorized as supervised learning. With Median Household Income as the target in our regression, the explanatory variables are those who were reported as being white, black, native american, asian and hispanic; along with percentage over 25 who completed high school, and the amount below poverty levels within each geographical area. If the results of the regression show statistically significant variables, then each can affect the median income, and each variable can be used to describe its relationship to said income. Our model is not intended for prediction necessarily, but to understand the relationship of these independent groups to median income; The estimated regression can be shown as:

$\mu\{Median\ Income\ |\boldsymbol{X}\} = -12883.73 - 945.64BELOWPOVERTY + 616.26COMPLETEDHS + 229.06WHITE + 254.47BLACK + 321.37NATIVE + 1502.76ASIAN + 260.37HISPANIC$

As seen in Figure 8, given there is evidence at a 95% confidence level that each coefficient is not zero, all groups can have a significant effect on the mean of median income. While each ethnic group is calculated to have positive estimated means, some are higher than others. For example, all else being equal, the group for those who identified as Asian on average have a median income that is $1242.39 more than the Hispanic group. For context, since the R-squared statistic measures the proportion of the total variation in median income around its mean that could be captured by the estimated regression, our model accounts for only 46% of the variation in median income, leaving 54% of the variation unaccounted for.

Figure 8

```
Call:
lm(formula = Median.Household.Income ~ X..below.poverty.level +
    X..over.25.completed.hs + share_white + share_black + share_native_american +
    share_asian + share_hispanic, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-78914  -9894  -3101   5713 172825

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -12883.73    3079.79  -4.183 2.89e-05 ***
X..below.poverty.level    -945.64      14.06 -67.240  < 2e-16 ***
X..over.25.completed.hs    616.26      16.64  37.037  < 2e-16 ***
share_white                229.06      27.54   8.317  < 2e-16 ***
share_black                254.47      29.04   8.763  < 2e-16 ***
share_native_american      321.37      30.40  10.573  < 2e-16 ***
share_asian               1502.76      48.15  31.207  < 2e-16 ***
share_hispanic             260.37      13.30  19.574  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18330 on 19101 degrees of freedom
Multiple R-squared:  0.4559,    Adjusted R-squared:  0.4557
F-statistic:  2287 on 7 and 19101 DF,  p-value: < 2.2e-16
```

Conclusion

The choice to portray our data using a linear regression was predicated around the multi-causal relationship that many of the factors had among each other. Initially -- utilizing partition clusters -- we had attempted to determine the correlation between race, education, income, and poverty levels; however, this approach proved inadequate as many clusters were insignificant to determine appropriate insight. Instead, a linear regression was sufficient to prove said variable correlation(s) true. The confidence interval determined that there is some relevance between the factors utilized in this regression; and, henceforth, face and education play a significant role in determining income.

Perhaps the explanation for a correlation in education in income can be best corroborated by Dennis Viloro from the U.S. Bureau of Labor Statistics who, after conducting an analysis by grouping workers by their respective education level, concludes that “More education leads to better prospects for earnings and employment” (Viloro 2016).

As a result, the proposed solution is to increase funding towards public education in low-income areas in order to reduce poverty levels. For further clarification, analyzing the gap in education in relation to race was necessary to determine two key impacts. First, income directly plays a role in poverty levels with education being one of the factors in determining income as depicted in Figure 7 with the percentage of people that completed high school (x-axis) and the respective income (y-axis). Second, the areas with a limited education were much easier to locate for allocation of funding, e.g. whether or not a supermajority of a city has a high school level education.

Our findings indicate that not only do different ethnic groups exhibit varying effects on the national mean of median income, but that cities which have underserved education standards are likely to continue having reduced levels of income. By using analytical tools, we are able to combat the prevailing bias of race and socioeconomic status by recommending public policy to level the education playing field, thus mitigating elements that continue to negatively influence communities.

References

Dennis Vilorio, "Education matters," *Career Outlook,* U.S. Bureau of Labor Statistics, March 2016.

Noël, Reginald A. “Race, Economics, And Social Status” bls.gov/spotlight/2018/race-economics-and-social-status/pdf/race-economics-and-social-status.pdf. 2018. Print.