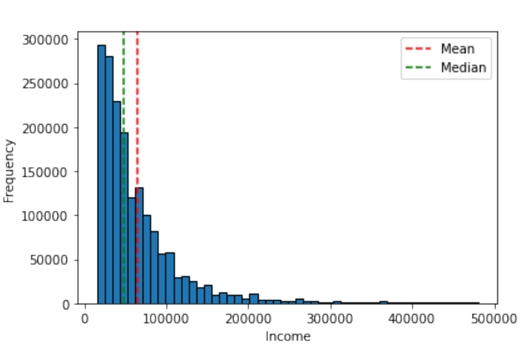
Tom Cooklin

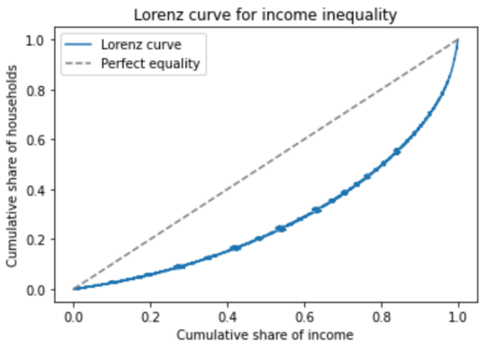
Final Report

Introduction

Income inequality is a prevalent issue in American society. Income inequality is higher in the US than any other developed country. Income inequality in the US also rose in 2021 for the first time since 2011. Both of these statistics highlight an issue rooted in American society which needs urgent attention. The motive behind my project is to utilize as many methods and skills I have learnt in my undergraduate studies to explore the discrimination behind income, such as how the sex and race of an individual affects their income. I also intend to explore the relationship between the race composition and diversity of a community with income inequality and if the race/sex of an individual can be predicted given employment statistics. To approach this project, I selected to use the American Community Survey (ACS) 2021 Microdata. In this dataset, each row is an individual which contains geographic, demographic and economic data of the individual. The dataset has over 3 million observations and 500 variables.

Income in the US is heavily right skewed which was a common problem in my project as outliers skewed summary statistics and models. The majority of my project involved median calculations rather than mean to reduce the effect of outliers.

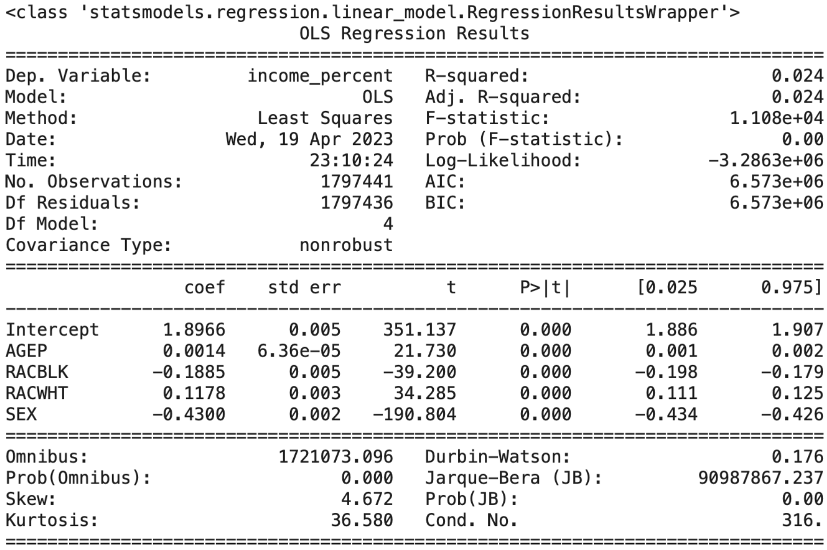




The ACS microdata provides Public Use Microdata Areas (PUMAs) which compromise of at least 100,000 individuals and are more geographically specific that states. To compare incomes across PUMAs, which have different costs of living, I divided each individual's income by the median income of their PUMA.

Individual Approach

Below shows the result of a linear model where RACBLK indicates a black individual, RACWHT indicates a white individual and AGEP is the age of the individual. My target variable was "income_percent" which was calculated using the methodology already described.

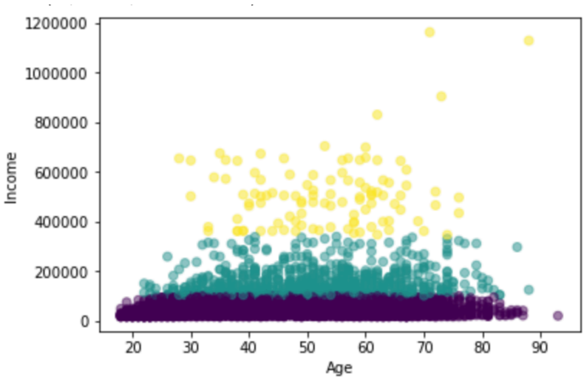


```
<class 'statsmodels.regression.linear_model.RegressionResultsWrapper'>
                            OLS Regression Results
==============================================================================
Dep. Variable:         income_percent   R-squared:                       0.024
Model:                            OLS   Adj. R-squared:                  0.024
Method:                 Least Squares   F-statistic:                 1.108e+04
Date:                Wed, 19 Apr 2023   Prob (F-statistic):               0.00
Time:                        23:10:24   Log-Likelihood:            -3.2863e+06
No. Observations:             1797441   AIC:                         6.573e+06
Df Residuals:                 1797436   BIC:                         6.573e+06
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.8966      0.005    351.137      0.000       1.886       1.907
AGEP           0.0014   6.36e-05     21.730      0.000       0.001       0.002
RACBLK        -0.1885      0.005    -39.200      0.000      -0.198      -0.179
RACWHT         0.1178      0.003     34.285      0.000       0.111       0.125
SEX           -0.4300      0.002   -190.804      0.000      -0.434      -0.426
==============================================================================
Omnibus:                  1721073.096   Durbin-Watson:                   0.176
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         90987867.237
Skew:                           4.672   Prob(JB):                         0.00
Kurtosis:                      36.580   Cond. No.                         316.
==============================================================================
```

The results show that white people and males earn more than black individuals and females. Additionally, the model indicates that an increase in age is associated with higher income which is as expected with more experience and qualifications.

Clustering

Next, I clustered the data into three different groups using the K-Means algorithm. The algorithm partitioned on income as seen below. The silhouette score for this clustering was 0.685 which can be considered relatively good. To avoid confusion, there are 4 different races in my dataset which include white, black, Asian and other races. Due to small sample sizes races outside of white, black and Asian were combined into another category called other.

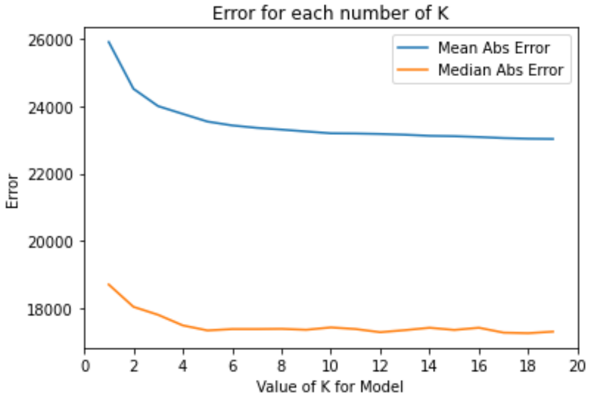


Analysis of these clusters yielded that Asian individuals appear the most in the highest income cluster, mainly due to them receiving the most degrees past undergraduate degrees which is valuable to entering this cluster. Black and other races were found most commonly in the lowest income cluster. Females are represented less than men in higher income clusters and more in lower income clusters despite females earning more college undergraduate degrees. This result is relevant because education shares the strongest correlation with income and females are represented unfairly despite earning more degrees. Females are also found more in better earning classes of work such as being self-employed or working for private for-profit companies. The receival of undergraduate degrees is consistent across all races. This could be caused by incentives and programs to include a diverse community in college undergrads. These programs tend to decrease for graduate level as programs are more selective, hence why we see a drop off in black and other races in post college undergraduate education. Black individuals are most commonly found in government jobs which earn the least. This could be caused by black and other races being considerably less likely to receive any college education.

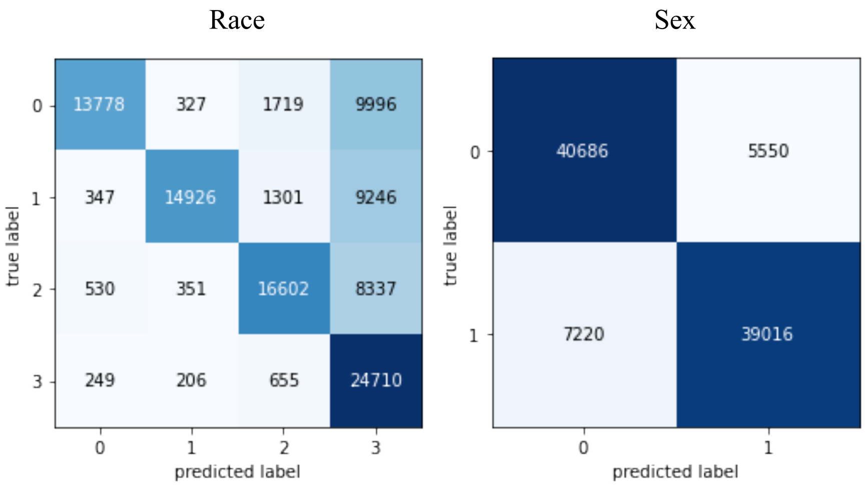
Holistically, using feature importance, highest level of education attained had the biggest impact on how much an individual earn. Despite this, when looking at individual clusters, the age of an individual was the most important factor. Highest level of education attained was the second most important in the lowest income cluster and became less important going through income clusters. This could imply that factors such as connections and experience play a larger role in attaining higher income jobs whilst education becomes less valuable. For the highest income cluster, the race of the individual is the second most important factor in determining income. Ideally, race should have no impact in deciding income and its importance in determining income in high income jobs shows the presence of systemic racism, whether that be in recruitment of education accessibility.

Predicting Race and Sex

Initially, I wanted to predict a person's income from their age, sex, race, state and employment. The idea was that an individual could input their data and compare their results to a similar employee but of different sex or race to see if their income is somewhat unfair. I used ensemble learning by taking the average of 2 different models, KNN and a random forest. I also tried these models independently and an XGBoost model but the ensemble model provided the highest accuracy which was still relatively poor. The variation of income is too large and dependent on too many factors such as experience and connections which aren't included in the ACS microdata.



Instead, I approached the somewhat converse of predicting race and sex from an individual's income, employment, state and age using a random forest classifier. The confusion matrices respectively are shown below.


Race

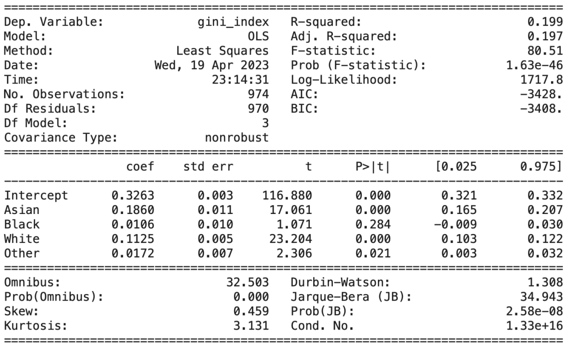
Sex

The test sample included an even distribution of race/sex so for race (if no bias was present) the model should be correct ~25% of the time and for sex, ~50% of the time. The accuracy score for predicting sex was 86.2% which was significantly higher using a chi-squared test. The accuracy score for predicting race was 67.8% which was also significant using a chi-squared test. The ability to, to some degree, predict someone's race or sex from their income and employment highlights the presence of racism and gender inequality in the US.

Community Approach

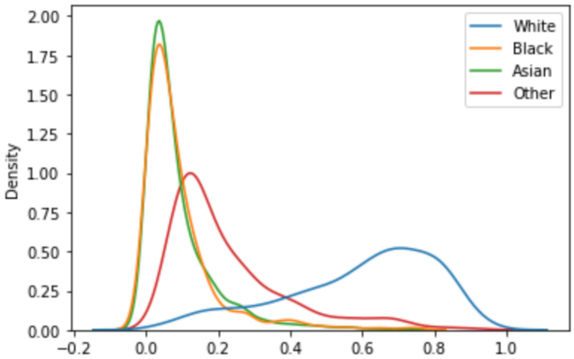
The Gini index is a measure of income inequality in a population. A Gini index score of 0 indicates perfect equality whilst a score of 1 indicates perfect inequality. Its calculation (see Lorenz graph on page 1) can be seen as the area between the Lorenz curve and line of perfect equality divided by the area underneath the line of perfect equality. Utilizing the Gini Index, the PUMAs Dade County in Miami had the best income equality whilst Highland Park Town in Dallas had the worst.

I wanted to see if there was any relationship between the presence of a race in a community and the Gini index of that community. I calculated the percentage of each race in each community and its respective Gini index and inputted this in a linear model with Gini index being the target variable.

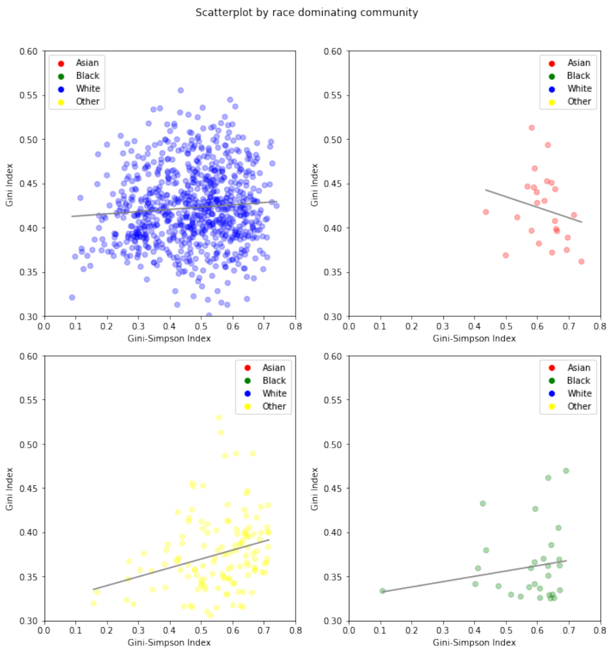


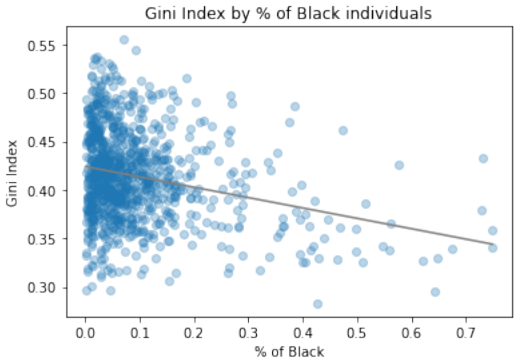
```
==============================================================================
Dep. Variable:             gini_index   R-squared:                       0.199
Model:                            OLS   Adj. R-squared:                  0.197
Method:                 Least Squares   F-statistic:                     80.51
Date:                Wed, 19 Apr 2023   Prob (F-statistic):           1.63e-46
Time:                        23:14:31   Log-Likelihood:                 1717.8
No. Observations:                 974   AIC:                            -3428.
Df Residuals:                     970   BIC:                            -3408.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.3263      0.003    116.880      0.000       0.321       0.332
Asian          0.1860      0.011     17.061      0.000       0.165       0.207
Black          0.0106      0.010      1.071      0.284      -0.009       0.030
White          0.1125      0.005     23.204      0.000       0.103       0.122
Other          0.0172      0.007      2.306      0.021       0.003       0.032
==============================================================================
Omnibus:                       32.503   Durbin-Watson:                   1.308
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.943
Skew:                           0.459   Prob(JB):                     2.58e-08
Kurtosis:                       3.131   Cond. No.                     1.33e+16
==============================================================================
```

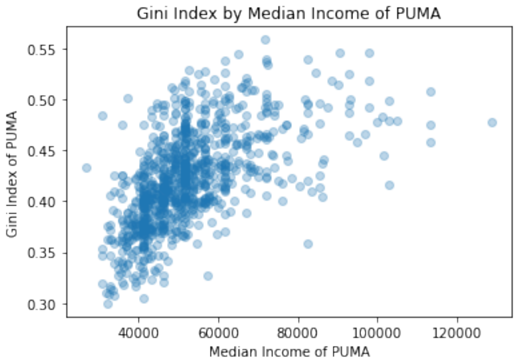
The results reveal no clear relationship between race composition and the Gini index of a community. There is also nothing to suggest there is a racial composition of a community which optimizes equality.



From my personal education, diverse communities are communities which thrive. I wanted to see if there was a relationship between the diversity and income inequality of a community. To measure diversity, I used the Gini-Simpson index which is the probability of picking 2 individuals from a population and them being of the same race. Despite educational thinking, there is nothing to suggest that the diversity leads to better income equality. I separated each color representing the dominant race in the community into independent scatterplots. Communities dominated by black and other races see an increase in income inequality as they become more diverse whilst white and Asian see the converse. Logical thinking from the individual approach would be that an increase in white and Asian individuals would lead to increase in income and income equality. The reasoning behind this relationship is quite the opposite. The Gini index takes into account the spread of income but not the average income of a community. Poorer communities have less spread of income and therefore what is considered better income equality. On average, black and other races earn less than white and Asian people so an increase in these races leads to a lower average income, less spread and a better Gini index score. The below plot shows the described relationship for black individuals and different communities. The same relationship is seen for other races and the converse for white and Asian individuals.







These results demonstrate that whilst the Gini index is good indicator of income equality, it is a poor indicator of income fairness. It doesn’t take into account whether a community is poor or wealthy. If I was to approach this project again, I would attempt to incorporate the average income into my models. Standardizing income using a normal distribution would remove the effect of average income disparities between communities but income follows a lognormal distribution and this would remove information valuable to my research. A better approach would be to utilize different indexes which account for average income such as the Atkinson or Theill index.

Conclusion

It is clear that systemic racism and gender equality exists in the US. Females earn less than males despite being more educated and working in prosperous classes of employment. Black and people of other races earn less than white and Asian individuals mainly due to lack of access to college education. From these biases, it is possible to predict the race and sex of an individual with a significant degree of accuracy. From a community perspective, diversity has no influence over the income equality of a community. An increase in black and other races in a community translates to better income equality highlighting the Gini indexes effectivity to measure income equality but not income fairness. Public officials should be wary of using the Gini index to make informed decisions.