Income Bias Across People with Different Demographic Backgrounds in New York

Baoling Zhou
New York University
Center for Urban Science and Progress
New York, NY, USA
bz882@nyu.edu

Shreya Bamne
New York University
Center for Urban Science and Progress
New York, NY, USA
ssb602@nyu.edu

Gerardo Rodríguez Vázquez
New York University
Center for Urban Science and Progress
New York, NY, USA
grv232@nyu.edu

Introduction

1. Problem context

LinkedIn is the social media platform for working professionals and is the best way to remain updated with latest topics in the professional world. LinkedIn sends a daily update comprising of latest news and one such article we came across was about racial discrimination in hiring people for jobs. Every job requires a different skill set and has salary associated with the level of expertise required for the same. Unfortunately, demographics are also affecting the income along with skill sets and this definitely needs understanding of the causes of the same. With this aim we decided to study income bias for the state of New York based on demographics. The effects of a range of different factors on the wages of Hispanics, Asians and blacks, relatively to whites and separately for women and men are examined. Significant differences are found in the sources of wage inequality across race, ethnicity, and gender.

2. Purpose of study

Understanding the income bias and more importantly its causes can help in addressing the issue. There are plenty of papers written on this issue and we are trying to understand if the income bias has remained the same or undergone some changes. This is important as there are have been drastic changes in the industry and economy. Recently, scholars have extended this approach in considering how black have fared as a result of changes in the racial, ethnic, and immigrant composition of the United States (McCall, 2001). Therefore, this study would help us understand the current scenario of income bias but with the similar set of variables, focusing on demographics. Another reason to focus on demographics is that previous studies have shown that demographics play an important role in income bias. We therefore decided to understand income bias using individual level census dataset of year 2016. We especially want to understand the income bias with respect to demographics.

3. Statement of responsibility:

We are a group of three and have worked collaboratively to complete this project. The individual contribution of each group member is as follows:
Baoling and Shreya have done data cleaning, coding for model and plotting results collaboratively. Gerardo has thought of initial idea of model and also searched for papers that have done similar research before. Shreya has written the report for Introduction, Literature review, Data and data cleaning and descriptive analysis. Baoling has written the report for Model, Results and Conclusion.

Literature Review

Income bias has been studied since a long time and these studies are based on various factors like industrial restructuring, demographics, employment conditions, etc. The common thing observed about these studies is that they have always included the demographics to understand income inequality. Due to this we have looked at particular studies that focus on income bias based on demographics. Another reason to take up this research is that we want to study income bias for the state of New York which has diverse demographics.

The first study we found was on Racial Wage Inequality (Leslie McCall, 2001) which talks about impact of race on wages by taking into consideration other demographics such as age and ethnicity. This study uses a model that considers race as binary variable and has some coefficients associated with each race. These coefficients were calculated using an equation comprising of some values calculated for gender and ethnicity of that race. The overall study indicates that there are various sources of income inequality and these vary for race, gender and ethnicity. Apart from demographics, the author has also discussed how immigration has impacted the wage inequality. In addition to this other significant takeaway from this study are that the income inequality varies spatially therefore there is a need to consider the spatial context as some ethnic group might be concentrated in certain region.

The other study we looked at was actually suggesting improvement to a model to understand income inequality. The original model used various independent variables like Years of school, Work Experience, Age, Marital status, Living in urban or rural area, number of children. The improvement made to the model was adding Cognitive Skill as another independent variable to the above mentioned list. They found that their model could account for 109% racial wage inequality whereas the previous model accounted for only 80%.(George Farkas, Keven Vicknair, 1996) The study suggests that using Cognitive Skill affected the results of the model significantly. This emphasizes on the sensitivity of models to variables and importance of variable selection in analysis on income bias.

Data Collection and Cleaning

We are using the IPUMS Census Data 2016 at individual granularity in New York City. It comprises of 284 columns and 196104 rows initially. It is a huge dataset comprising of lot of variables. Since our research is about understanding income bias, we decided to choose a group of variables which would help in addressing the problem.

As we are looking at income bias, we decided to select the 'WAGP" variable as our dependent variable which denotes wages or salary income for past 12 months. Our research aims at understanding income bias based on demographic features. Therefore, we selected various variables which include Age, Sex, Race, Education Attainment, Nativity, Citizenship, Marital Status, Children(gave birth to child in last 12 months) and job Industry.

Thus after selecting the variables we got rid of unwanted columns and proceeded with data cleaning.We began with filtering the dataset by considering each independent variable at a time, starting with Age variable. For Age group, we considered a range of 25 to 64 years. This age range was selected due to that industries of our interest had majority of people falling in this age range. Also, since we are planning to look at citizenship status, majority of the non-citizens fall in this age range. After filtering the data for the age ranging from 25 to 64, we were left with 102696 records.

The next variable we considered was "Race". The dataset comprised of nine different races. We found that White, Black, Asian and Hispanic are the majority. Therefore we filtered the dataset by considering individual records that belonged to these four categories and we were left with 100108 records.

Education attainment is another important variable due to its strong correlation with income. The dataset has 25 different levels of education attainment, starting with Preschool to Doctorate. There was definitely a need to filter on this variable addressing to the selected age range. We decided to take education attainment from "Regular High School Diploma" to "Doctorate". Also, this level of education complies with the industry we have chosen and also has non-US citizen group included. On applying this filter we ended up with 89452 records.

Whenever we speak of income, an important thing to consider is the job industry. The dataset has 267 different categories which can be grouped into 18 broad categories, out of which we chose 7 broad categories. The dataset was filtered again with respect to Industry categories of Finance, Education, Manufacturing, Medical, Administration, Information and multiple professions. With this, we ended up getting 44441 records in our final dataset which was then used for modelling.

The dataset comprised of categorical variables and many of them were in the tabulated form. However, a few variables were as per codes used by IPUMs. We therefore, converted these variables into categorical variables. The changes were made for Marital Status, Education Attainment, Citizenship and Industry variables. For Marital Status, we had five different categories which were converted into two i.e. Married and Unmarried (separated, widowed,

divorced, never married). Education Attainment had nine categories which were reduced to three categories i.e. Less than Bachelor degree, Bachelor degree and Bachelor degree and above. Citizenship had four categories which were reduced to two i.e US citizen and not US citizen. For industry we created dummy variables for each type, therefore seven dummy variables in total. With this we ended up with a dataset of 44441 rows and 24 columns.

Descriptive Analysis

After having cleaned the data, we wanted to get the summary of data in order to understand it well before proceeding with the model. To do this, we saw how income varied with respect to average income for different demographic variables, especially for categorical variables as we can easily visualize them and get some useful insights for further analysis.

Exhibit 1 shows visualizations for average income comparison regarding different demographics. We find that for citizenship status, US Citizens have income equal to average income as compared to Non-US Citizens. On comparing income by marital status, the married people are observed to have more than average income. Looking at race variable, whites and asians are having more than average income whereas other two races are below average range. Comparing by gender shows that men have more than average income while females are below the average line. These are of course some preliminary observations based on the cleaned dataset. We would definitely get better insights on why the results are the way they are after performing multivariate regression analysis which is explained in the next section.

Methods: Multivariate Regression Analysis

Regression model: $WAGP = \beta 1\, AGEP + \beta 2\, C(SEX) + \beta 3\, C(CIT) + \beta 4\, C(MAR) + \beta 5\, C(FFERP) + \beta 6\, C(race) + \beta 7\, C(Educ) + \beta 8\, C(FIN) + \beta 9\, C(EDU) + \beta 10\, C(MED) + \beta 11\, C(PRF) + \beta 12\, C(MFG) + \beta 13\, C(ADM) + \beta 14\, C(INF) + \alpha + \varepsilon$

***Dependent Variable*:**

- **WAGP**: Wages or salary income past 12 months ($1 to 999999)

Independent Variable:

- **AGEP**: numerical value - Select age group 25 to 64
- **SEX**:sex
    - 0: Female
    - 1: Male

- **race**: race code
    - 0: White
    - ‘Asian’: Asian alone
    - ‘Black: Black or African American alone,
    - ‘Hispanic’: Spanish/Hispanic/Latino
- **Educ**: Educational attainment
    - 0. Less than Bachelor Degree
    - 1: ‘Bachelor’. Bachelor Degree
    - 2: ‘Master’. More than Bachelor Degree
- **NATIVITY**: Nativity
    - 1: Native
    - 2: Foreign Born
- **CIT**: Citizenship status
    - 0: Not a citizen of the U.S.
    - 1: U.S. citizen
- **MAR**: Marital Status
    - 0: Not married
    - 1: Married
- **FFERP:** Gave birth to child within the past 12 months allocation flag
    - 0: No
    - 1: Yes
- **Industry (binary variable):** we select the following industries because there are usually education barrier to get into these industries.
    - FIN: Finance, eg. Banking, real estate, financial investment
    - EDU: Education, eg. Elementary schools, college
    - MED: Medical, eg. hospitals, dentists
    - PRF: multiple professions, eg. Accounting, Computer science, Engineering
    - MFG: Manufacturing, eg. Industrial and chemicals
    - ADM: Administration, eg. Human resources, city agencies
    - INF: Information, eg. publishers, telecoms, data processing

The alpha has been set as 5%. Then, a step-backward feature selection is utilized to further refine the model. The step-backward method is carried out by continuously removing repressors (one by one) having the lowest t-stats, until the R-squared value can no longer be improved. This will allow for a stronger relationship within the model, and create a better predictive measure. We excluded nativity from our model since it has multicollinearity with citizenship binary variable.

Results

Exhibit 2 show the output result of the final regression model. The R-square is low which is only 0.178 (Exhibit 2). Considering the scope of study is social science, we say the prediction power of the model is acceptable.

The coefficient of all variables are significant at 95% confidence level except for 'Asian' and FFERP (giving birth to child within past 12 months). This shows that giving birth to child within a year or being Asian actually won't make a person earn different level of income compared to people with same backgrounds and working in the same industry. We ended up remaining these two variables in the model since R-Squared would decrease if they got eliminated.

We see the following income bias based on the regression model we built. Holding all other variables constant, males can earn $28657 more in annual income than female do on average. This shows a potential income bias caused by gender discrimination. However, since don't have each person's job title, we cannot exclude the possibility that male usually work at higher positions than female do in a certain industry, which can cause them earn different level of income. From another perspective, this result may prove the existence of an unfair working environment for female employees as there are less opportunities for them to get promoted. In addition, the coefficient of citizenship binary variable shows us that having US citizenship will make a person earn $4598 in annual income than someone who does not, holding all other factors the same. This can be explained by the cost of maintaining foreign employee's visas from the perspective of companies (Exhibit 3).

Considering the impact of races on someone's annual income, it shows that a Black employee would earn $8109 less on average than a White employee does in the context that all other aspects are the same. This discrepancy turns out to be $11595 between Whites and Hispanics, which shows a more severe discrimination issue. This can be justified by the fact that Whites people usually dominate the high-level positions in a firm and shows an unbalanced human resources allocation (Exhibit 3).

On the other hand, it is noted that married people usually earn $14133 more annually than unmarried people do on average, holding all other factors the same. This can be explained by a more stable lifestyle and better reputation caused by the marital status. In addition, it is not surprising to see that people generally earn higher income as they get older. It can be explained by the year of experience and getting promoted to higher positions. More specifically, one age increase can let people earn $709 more in annual income on average, holding all other factors the same. Regarding education attainment, people having bachelor degrees and the ones having master degrees can earn $34942 and $67133 more respectively than people who only have high school diplomas. This can be explained by the scarcity of the skills and knowledge.

Limitation

Even though we got significant coefficients for majority of the variables, it is not deniable that the prediction power of the model is very low. There are a lot of other factors can affect the annual income a person is earning, such as, the level of position, the size of firm and years of experience of the person (the age doesn't necessarily say it). Besides, there might be multicollinearity existing among different independent variables, such as age and marital status. For this reason, the coefficient result noted from the regression analysis could be biased. Lastly, the results are exclusively for New York state in 2016. We won't be able to see the changes occurring over time and perhaps the income bias is decreasing year by year.

However, building this model is a good start regarding our purpose of study and it can be easily apply to data of other states and other years.

Conclusion

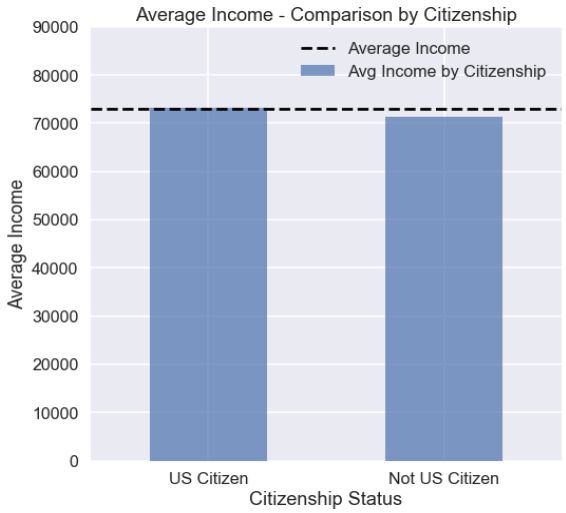
According to our regression analysis, we detect the existence of income bias caused by different genders and races for people who have similar background and work in the same industry. In addition, people who are married or have US citizenship possess advantages in earning higher income, but these have little to do with discrimination and is not in the scope of our study. On the other hand, our analysis shows that giving birth to child within past 12 months or being Asian actually won't make a big difference to someone's annual income, due to the insignificance results in coefficients. Asians have surpassed Latinos as the main immigrant group in the U.S. and their average income has been increasing over the years. This can be explained by the fact that well educated Asians have been entering the U.S. to pursue graduate degrees to later become part of the workforce.

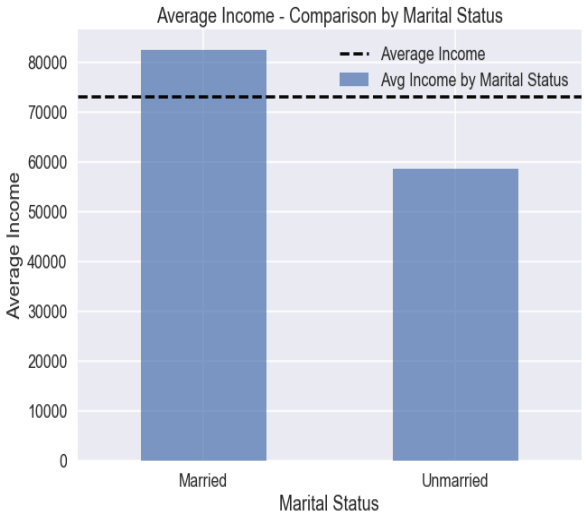
As for blacks and Latinos, our results echo previous theories that predict white managerial or elite exploitation of these communities as a source of cheap labor. They provide some justification for linking immigration to discriminations against the racial-ethnic groups most likely to be composed of immigrants. Also, immigrants tend to settle in the labor markets to their detriment and to the relative benefit of the white workers and managers.

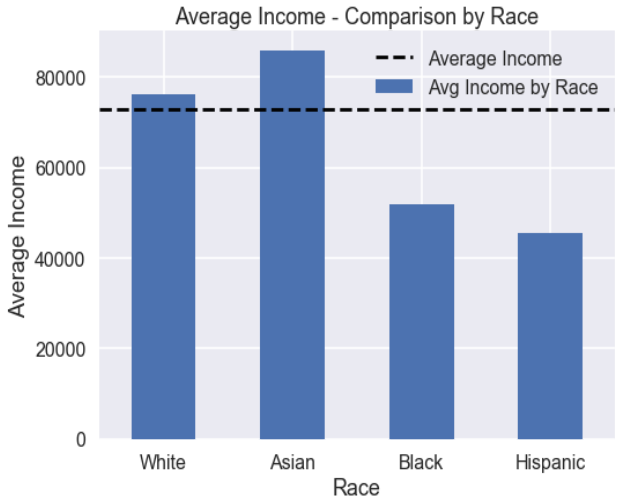
However, the above conclusion can be biased due to the low predicting power of the model, multicollinearity among independent variables and limited data observations. More research needs to be conducted to take more years and geographical areas into account and eliminate multicollinearity by doing Principle of Component Analysis. Overall, the regression analysis is a good start to explore the potential income bias and a more sophisticated model should be built for the purpose of finding reasons behind the results.

Exhibits

Exhibit 1: Average Income Comparisons for - Citizenship Status, Marital Status, Race and Gender







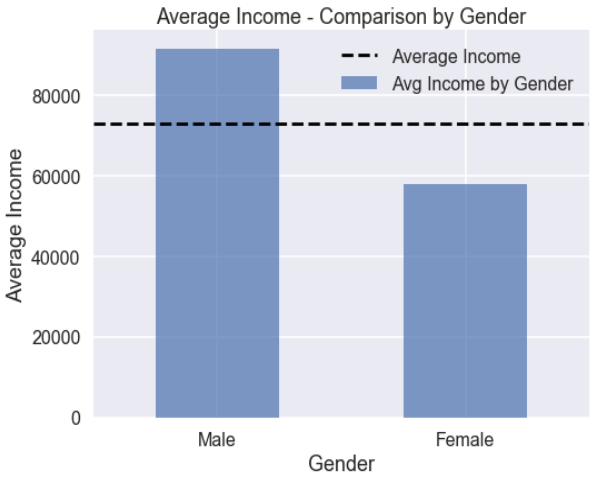


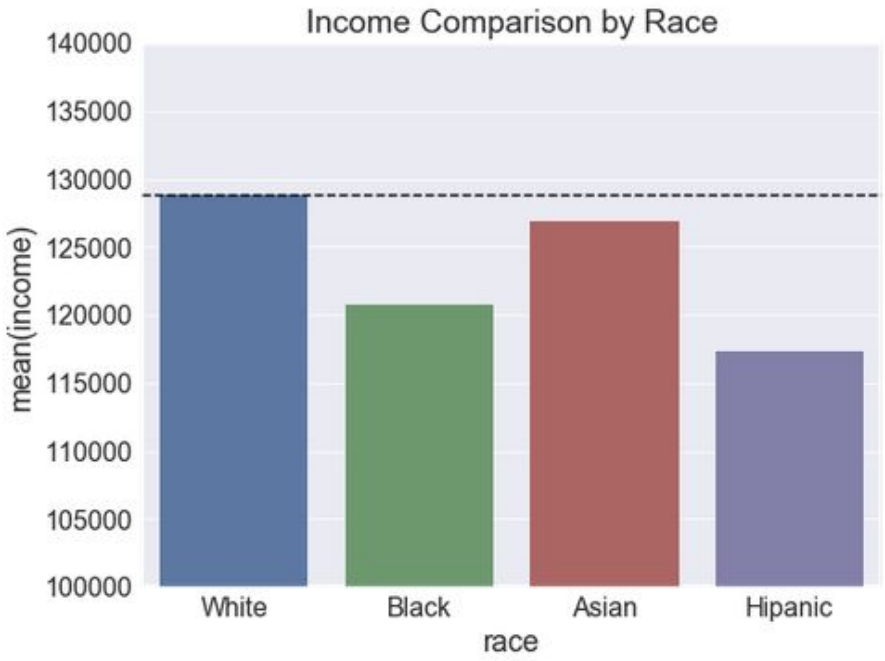
Exhibit 2: Final Regression Model Output (Alpha = 0.05)

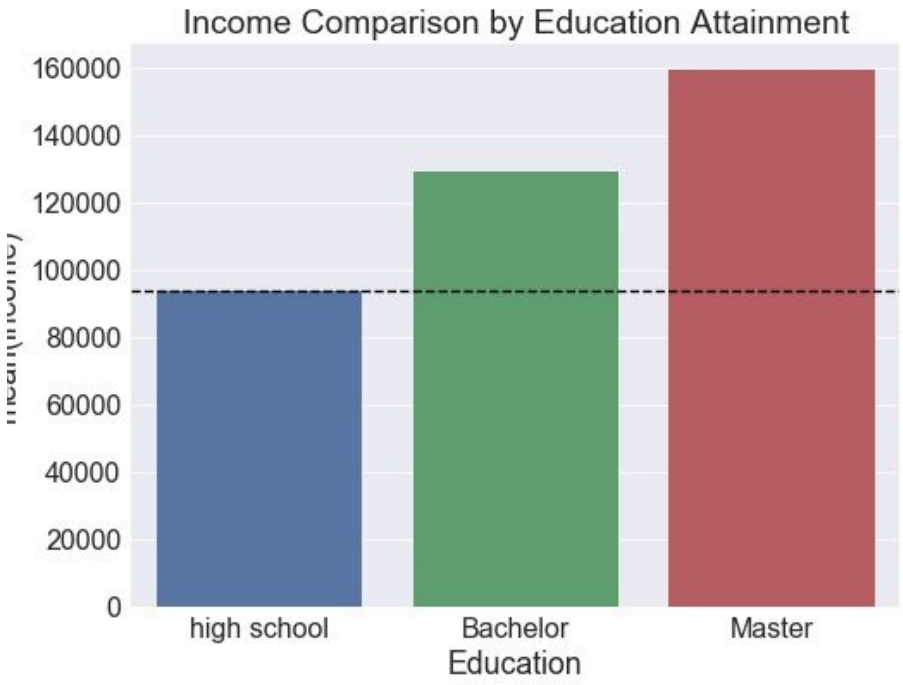
Dep. Variable:	WAGP	**R-squared:**	0.178
Model:	OLS	**Adj. R-squared:**	0.178
Method:	Least Squares	**F-statistic:**	601.2
Date:	Sat, 09 Dec 2017	**Prob (F-statistic):**	0.00
Time:	13:19:21	**Log-Likelihood:**	-5.6496e+05
No. Observations:	44441	**AIC:**	1.130e+06
Df Residuals:	44424	**BIC:**	1.130e+06
Df Model:	16		
Covariance Type:	nonrobust		

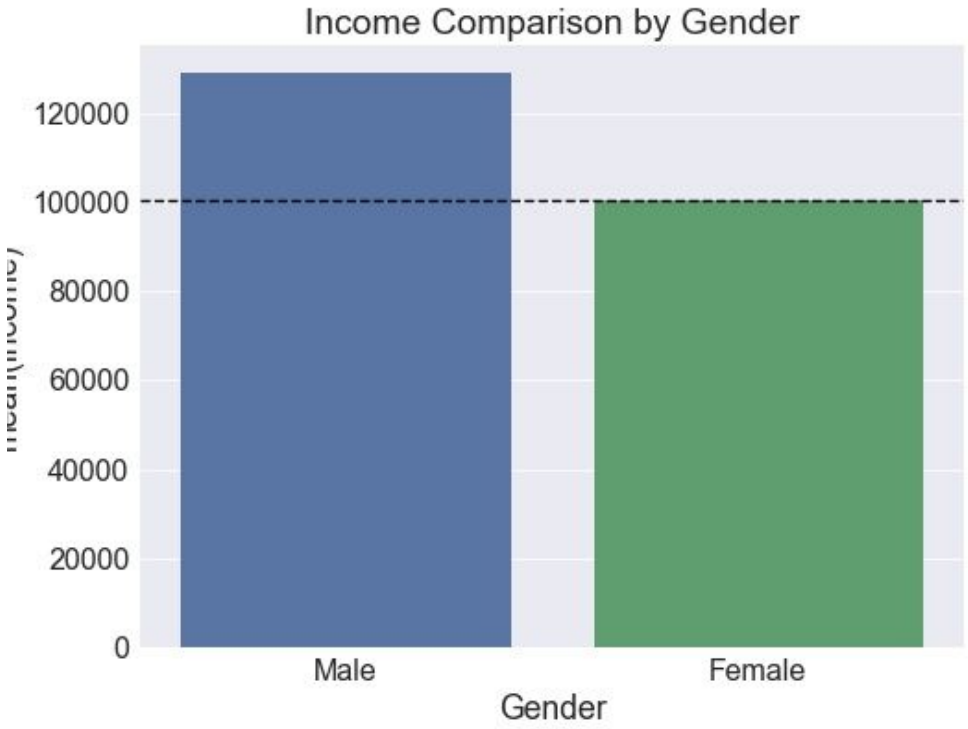
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8175.0375	2063.986	-3.961	0.000	-1.22e+04	-4129.589
C(SEX)[T.1]	2.866e+04	817.100	35.072	0.000	2.71e+04	3.03e+04
C(CIT)[T.1]	4598.7888	1612.402	2.852	0.004	1438.453	7759.125
C(MAR)[T.1]	1.413e+04	818.828	17.260	0.000	1.25e+04	1.57e+04
C(FFERP)[T.1]	3787.2173	2434.894	1.555	0.120	-985.218	8559.653
C(race)[T.Asian]	-2006.5429	1440.821	-1.393	0.164	-4830.577	817.491
C(race)[T.Black]	-8109.9713	1223.439	-6.629	0.000	-1.05e+04	-5712.010
C(race)[T.Hispanic]	-1.16e+04	2139.861	-5.419	0.000	-1.58e+04	-7401.049
C(Educ)[T.Bachelor]	3.494e+04	953.143	36.660	0.000	3.31e+04	3.68e+04
C(Educ)[T.Master]	6.713e+04	993.444	67.577	0.000	6.52e+04	6.91e+04
C(FIN)[T.1]	3.345e+04	1019.849	32.798	0.000	3.14e+04	3.54e+04
C(EDU)[T.1]	-3.098e+04	939.528	-32.978	0.000	-3.28e+04	-2.91e+04
C(MED)[T.1]	-3289.2369	871.230	-3.775	0.000	-4996.862	-1581.612
C(PRF)[T.1]	-2359.4892	879.968	-2.681	0.007	-4084.242	-634.736
C(MFG)[T.1]	-7161.2755	1127.623	-6.351	0.000	-9371.436	-4951.116
C(ADM)[T.1]	-3582.0437	1210.734	-2.959	0.003	-5955.103	-1208.985
C(INF)[T.1]	5752.4123	1566.003	3.673	0.000	2683.020	8821.805
AGEP	709.4746	35.821	19.806	0.000	639.265	779.684

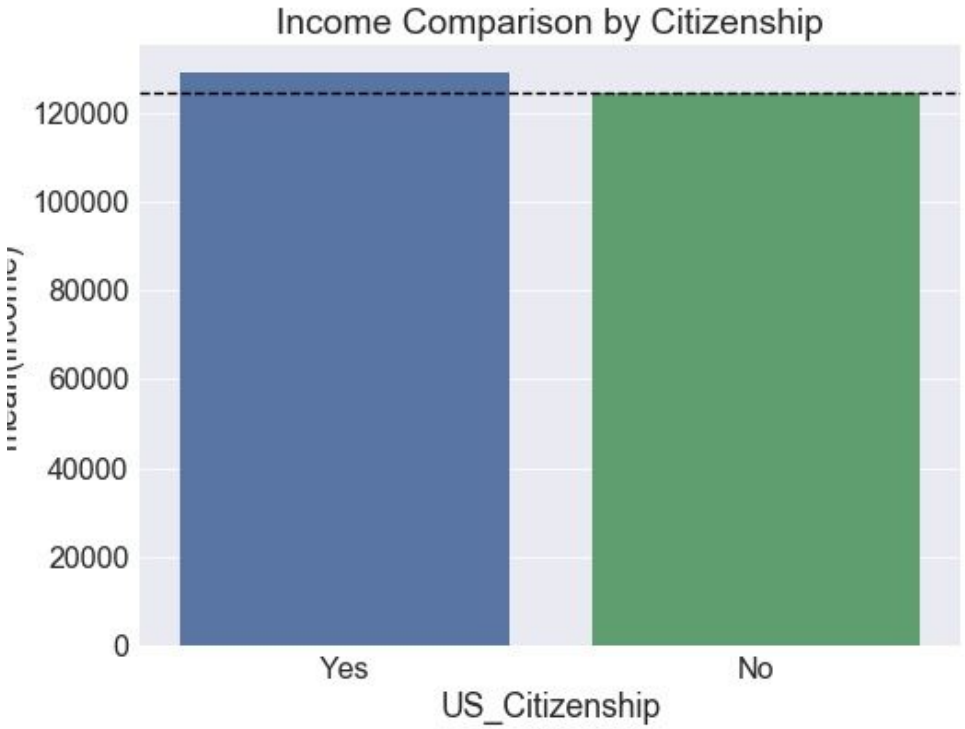
```
Intercept              -8175.037529
C(SEX)[T.1]            28657.008703
C(CIT)[T.1]             4598.788750
C(MAR)[T.1]            14133.170165
C(FFERP)[T.1]           3787.217304
C(race)[T.Asian]       -2006.542949
C(race)[T.Black]       -8109.971345
C(race)[T.Hispanic]   -11595.214664
C(Educ)[T.Bachelor]    34942.260727
C(Educ)[T.Master]      67133.978129
C(FIN)[T.1]            33448.725990
C(EDU)[T.1]           -30984.130466
C(MED)[T.1]            -3289.236900
C(PRF)[T.1]            -2359.489215
C(MFG)[T.1]            -7161.275524
C(ADM)[T.1]            -3582.043683
C(INF)[T.1]             5752.412268
AGEP                     709.474553
```

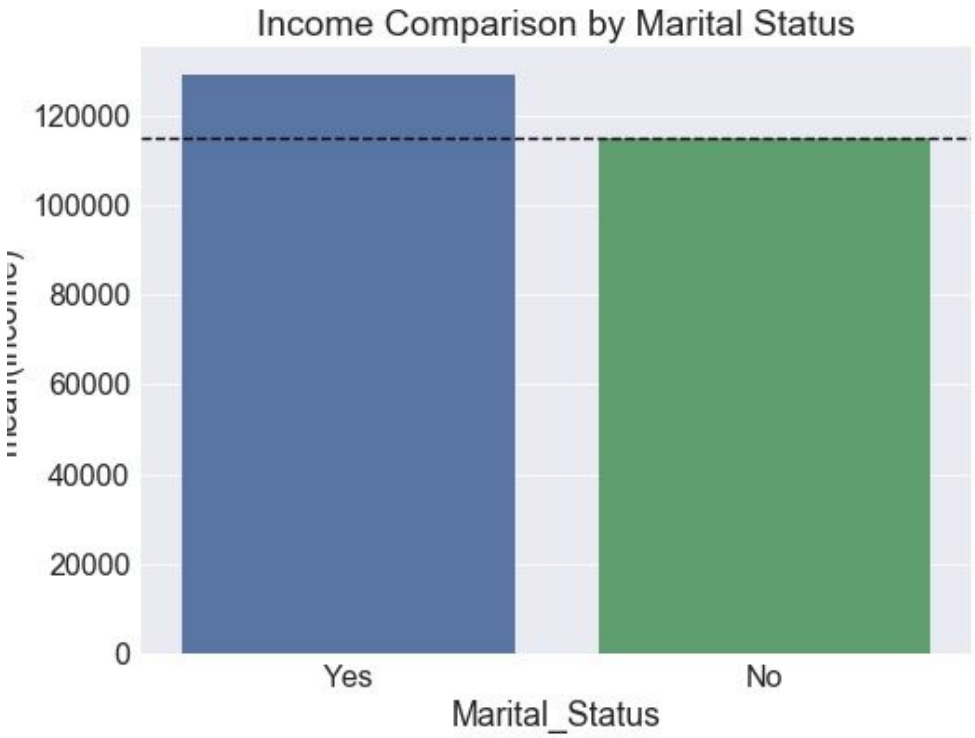
Exhibit 3: Income Bias caused by Race, Educational Attainment, Gender, Age, Marital status, Whether Having Citizenship, Whether gave birth in past 12 month, working industry.
We set default as 30-year old married White man with no child in past 12 month, having US citizenship and working in finance industry. Then, we made little adjustments based on default setting to generate the following comparison charts.

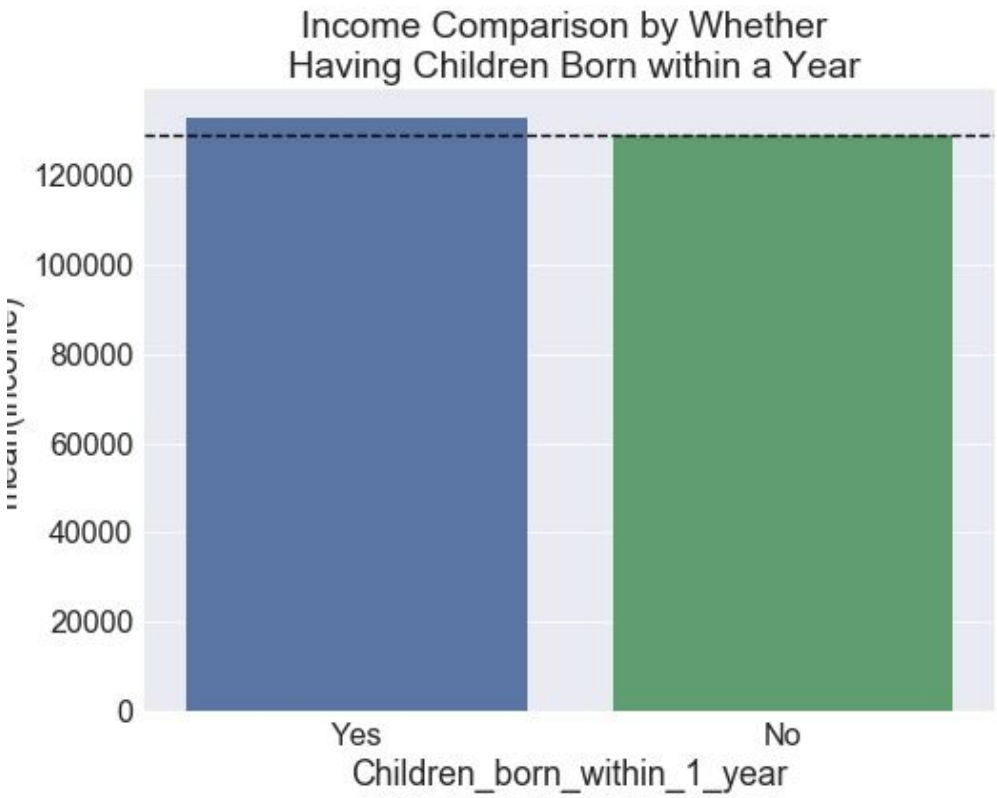


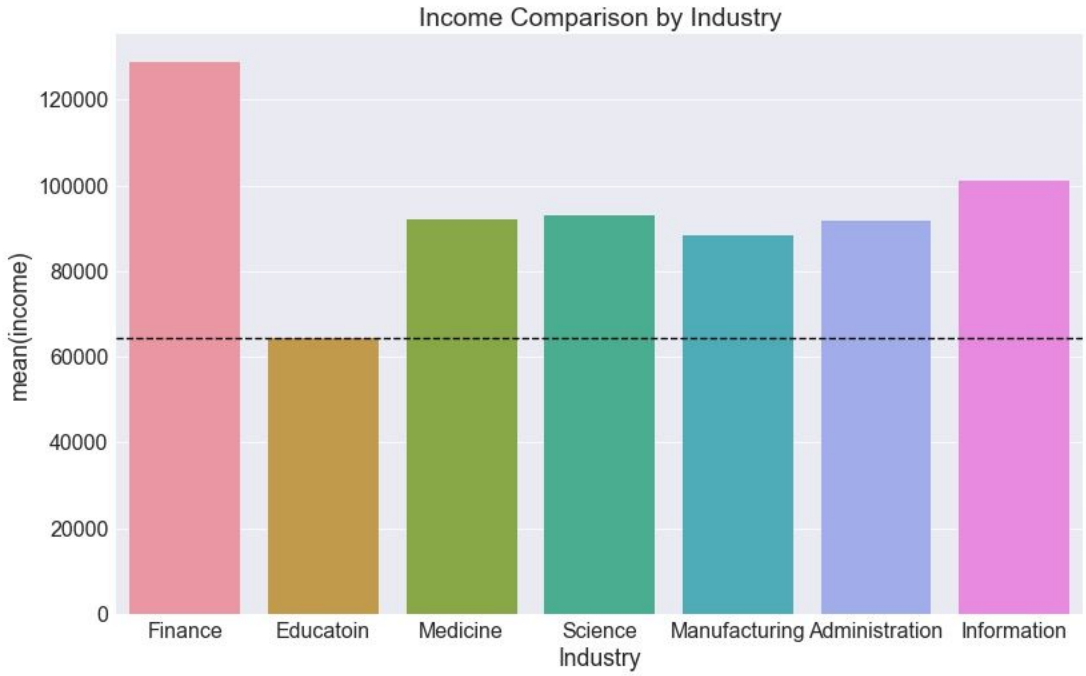

Income Comparison by Education Attainment
160000
140000
120000
100000
80000
60000
40000
20000
0
high school
Bachelor
Master
Education

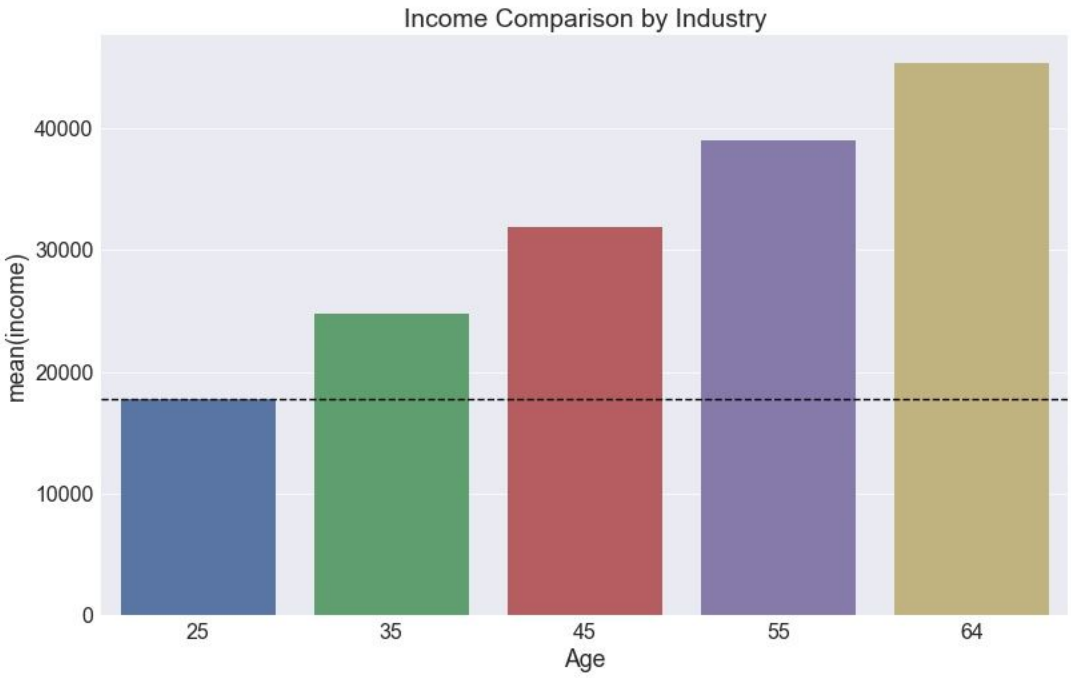

Income Comparison by Gender
120000
100000
80000
60000
40000
20000
0
Male
Female
Gender


Income Comparison by Citizenship
120000
100000
80000
60000
40000
20000
0
Yes
No
US_Citizenship


Income Comparison by Marital Status
120000
100000
80000
60000
40000
20000
0
Yes
No
Marital_Status


Income Comparison by Whether
Having Children Born within a Year
120000
100000
80000
60000
40000
20000
0
Yes
No
Children_born_within_1_year


Income Comparison by Industry
mean(income)
120000
100000
80000
60000
40000
20000
0
Finance
Educatoin
Medicine
Science
Manufacturing
Administration
Information
Industry



Reference

Leslie McCall, *Sources of Racial Wage Inequality in Metropolitan Labor Markets: Racial, Ethnic, and Gender Differences,* American Sociological Review, Vol. 66, No. 4 (Aug., 2001), pp. 520-541, http://www.jstor.org/stable/3088921

George Farkas, Keven Vicknair, *Appropriate Tests of Racial Wage Discrimination Require Controls for Cognitive Skill:Comment on Cancio, Evans, and Maume, American Sociological Review*, Vol. 61, No. 4 (Aug., 1996), pp. 557-560, http://www.jstor.org/stable/pdf/2096392.pdf

American FactFinder, 2016 ACS 1-year PUMS New York state Population Records. Retrieved December 09, 2017 from: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_pums_csv_2016&prodType=document

2016 ACS PUMPS Data Dictionary.(October 19,2017) US. census Bureau. Retrieved December 09, 2017[1] Focus on Poverty in New York City. (n.d.). Retrieved October 02, 2017, from: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_pums_csv_2016&prodType=document