ILLINOIS INSTITUTE OF TECHNOLOGY

# A Statistical Analysis of US Income by Education Level, Gender, Race, and Age

MATH 484/564 - Prof. Lulu Kang

Emily Piszczek and Thi Truong

12-02-2022

## Abstract

This project aims to better understand US income levels and its statistical relationship with earners' education level, gender, race, and age. Data from the 2020 US Census (*Educational Attainment-People 18 Years Old and Over, by Total Money Earning, Work Experience, Age, Race, Hispanic Origin, and Sex*) was used to create a weighted multilinear regression model. The census data was presented in mean earnings by group, where each group consisted of a specified education level, gender, race, and age. Alongside the mean earning by group, the data provided by the census also included sample size and standard error per group, all of which were included to create the MLR model from dummy variables using R software. Additional R tools and plotting methods were used to further analyze the relationships among the independent and dependent variables. The model showed a significant relationship for mean salary earnings and majority of independent variables, but not all. Although relationships show some groups earn a higher mean salary than other groups, correlation does not imply causation. One glaring limitation to the data was lack of location, since mean salary expectations vary around the US. The MLR model was determined to be mostly significantly significant, though analysis of the model shows imperfections and room for improvement.

## Introduction

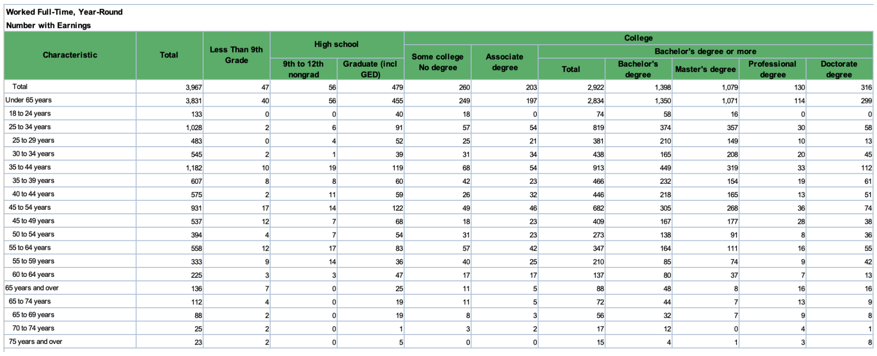
Pay equity in the US should result in salary earnings which reflect ability and experience, and should not discriminate based on gender or race. A statistical analysis is one way to examine if earnings are fair across different groupings of workers. It is also interesting to see how ability and experience, such as level of education, may impact salary earnings. This project examines demographic data taken from the 2020 US census to better understand how a few select variables impact salary earnings of full-time workers.

Because the census data was taken from across the US, there is likely a skew in result interpretation based on: cost of living by city, population concentrations across cities, and demographics across cities. Therefore this data analysis is not expansive enough to conclude if discrimination is present in the salaries of the US workforce. This project does not claim to draw such conclusions, but rather serve as a starting point for how age, gender, race, and education level may or may not impact expected earnings. The relationship between variables will be represented as a multilinear model. The analysis will include a variety of other statistical methodologies and models, and discuss their associated strengths and weaknesses.

## Data Collection

The primary source of data for this project was taken from the US Census Bureau's "Current Population Survey, 2021 Annual Social and Economic Supplement." The following values are based on adults who worked full time, year-round. The dependent variable being analyzed was the mean earnings for each grouping. Groupings consisted of the following four variables: education level, gender, race, and age. For this project, some of the categories were further simplified. For education level, only high school graduates (and GED), associate degree, and bachelor's degree or more were considered. This project did not consider those who did not complete high school or those with only some college experience. For gender, only males and females were considered. For race, only

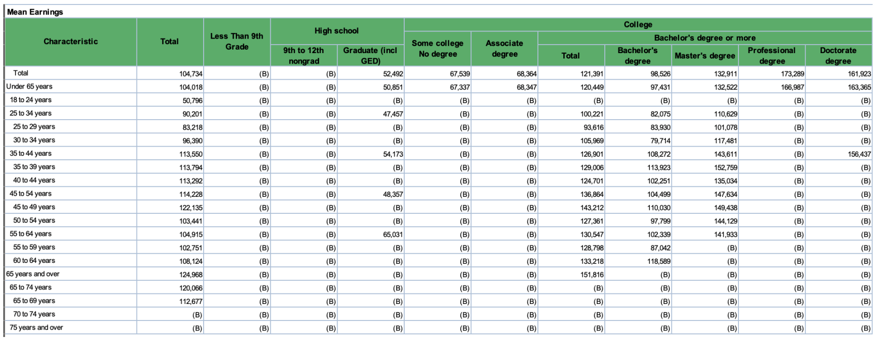
Asian, Black, Hispanic, and White categories were considered. Finally, age was broken up into seven categories: 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, and 75 and older. Below is an example of the sample size of Asian men (in the 1,000's) and the corresponding tables of their mean earnings and standard error of mean [2]:



Worked Full-Time, Year-Round
Number with Earnings

| Characteristic | Total | Less Than 9th Grade | High school | | College | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Some college No degree | Associate degree | Bachelor's degree or more | | | | |
| | | | 9th to 12th nongrad | Graduate (incl GED) | | | Total | Bachelor's degree | Master's degree | Professional degree | Doctorate degree |
| Total | 3,967 | 47 | 56 | 479 | 260 | 203 | 2,922 | 1,398 | 1,079 | 130 | 316 |
| Under 65 years | 3,831 | 40 | 56 | 455 | 249 | 197 | 2,834 | 1,350 | 1,071 | 114 | 299 |
| 18 to 24 years | 133 | 0 | 0 | 40 | 18 | 0 | 74 | 58 | 16 | 0 | 0 |
| 25 to 34 years | 1,028 | 2 | 6 | 91 | 57 | 54 | 819 | 374 | 357 | 30 | 58 |
| 25 to 29 years | 483 | 0 | 4 | 52 | 25 | 21 | 381 | 210 | 149 | 10 | 13 |
| 30 to 34 years | 545 | 2 | 1 | 39 | 31 | 34 | 438 | 165 | 208 | 20 | 45 |
| 35 to 44 years | 1,182 | 10 | 19 | 119 | 68 | 54 | 913 | 449 | 319 | 33 | 112 |
| 35 to 39 years | 607 | 8 | 8 | 60 | 42 | 23 | 466 | 232 | 154 | 19 | 61 |
| 40 to 44 years | 575 | 2 | 11 | 59 | 26 | 32 | 446 | 218 | 165 | 13 | 51 |
| 45 to 54 years | 931 | 17 | 14 | 122 | 49 | 46 | 682 | 305 | 268 | 36 | 74 |
| 45 to 49 years | 537 | 12 | 7 | 68 | 18 | 23 | 409 | 167 | 177 | 28 | 38 |
| 50 to 54 years | 394 | 4 | 7 | 54 | 31 | 23 | 273 | 138 | 91 | 8 | 36 |
| 55 to 64 years | 558 | 12 | 17 | 83 | 57 | 42 | 347 | 164 | 111 | 16 | 55 |
| 55 to 59 years | 333 | 9 | 14 | 36 | 40 | 25 | 210 | 85 | 74 | 9 | 42 |
| 60 to 64 years | 225 | 3 | 3 | 47 | 17 | 17 | 137 | 80 | 37 | 7 | 13 |
| 65 years and over | 136 | 7 | 0 | 25 | 11 | 5 | 88 | 48 | 8 | 16 | 16 |
| 65 to 74 years | 112 | 4 | 0 | 19 | 11 | 5 | 72 | 44 | 7 | 13 | 9 |
| 65 to 69 years | 88 | 2 | 0 | 19 | 8 | 3 | 56 | 32 | 7 | 9 | 8 |
| 70 to 74 years | 25 | 2 | 0 | 1 | 3 | 2 | 17 | 12 | 0 | 4 | 1 |
| 75 years and over | 23 | 2 | 0 | 5 | 0 | 0 | 15 | 4 | 1 | 3 | 8 |

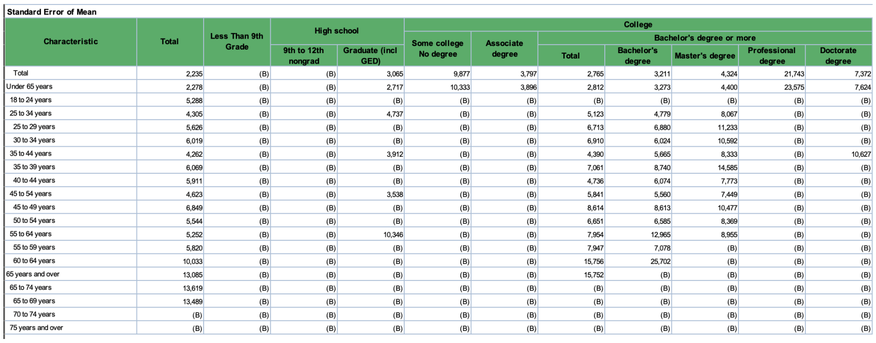
Figure 1: Sample sizes of Asian Male



Mean Earnings

| Characteristic | Total | Less Than 9th Grade | High school | | College | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Some college No degree | Associate degree | Bachelor's degree or more | | | | |
| | | | 9th to 12th nongrad | Graduate (incl GED) | | | Total | Bachelor's degree | Master's degree | Professional degree | Doctorate degree |
| Total | 104,734 | (B) | (B) | 52,492 | 67,539 | 68,364 | 121,391 | 98,526 | 132,911 | 173,289 | 161,923 |
| Under 65 years | 104,018 | (B) | (B) | 50,851 | 67,337 | 68,347 | 120,449 | 97,431 | 132,522 | 166,987 | 163,365 |
| 18 to 24 years | 50,796 | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |
| 25 to 34 years | 90,201 | (B) | (B) | 47,457 | (B) | (B) | 100,221 | 82,075 | 110,629 | (B) | (B) |
| 25 to 29 years | 83,218 | (B) | (B) | (B) | (B) | (B) | 93,616 | 83,930 | 101,078 | (B) | (B) |
| 30 to 34 years | 96,390 | (B) | (B) | (B) | (B) | (B) | 105,969 | 79,714 | 117,481 | (B) | (B) |
| 35 to 44 years | 113,550 | (B) | (B) | 54,173 | (B) | (B) | 126,901 | 108,272 | 143,611 | (B) | 156,437 |
| 35 to 39 years | 113,794 | (B) | (B) | (B) | (B) | (B) | 129,006 | 113,923 | 152,759 | (B) | (B) |
| 40 to 44 years | 113,292 | (B) | (B) | (B) | (B) | (B) | 124,701 | 102,251 | 135,034 | (B) | (B) |
| 45 to 54 years | 114,228 | (B) | (B) | 48,357 | (B) | (B) | 136,864 | 104,499 | 147,634 | (B) | (B) |
| 45 to 49 years | 122,135 | (B) | (B) | (B) | (B) | (B) | 143,212 | 110,030 | 149,438 | (B) | (B) |
| 50 to 54 years | 103,441 | (B) | (B) | (B) | (B) | (B) | 127,361 | 97,799 | 144,129 | (B) | (B) |
| 55 to 64 years | 104,915 | (B) | (B) | 65,031 | (B) | (B) | 130,547 | 102,339 | 141,933 | (B) | (B) |
| 55 to 59 years | 102,751 | (B) | (B) | (B) | (B) | (B) | 128,798 | 87,042 | (B) | (B) | (B) |
| 60 to 64 years | 108,124 | (B) | (B) | (B) | (B) | (B) | 133,218 | 118,589 | (B) | (B) | (B) |
| 65 years and over | 124,968 | (B) | (B) | (B) | (B) | (B) | 151,816 | (B) | (B) | (B) | (B) |
| 65 to 74 years | 120,066 | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |
| 65 to 69 years | 112,677 | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |
| 70 to 74 years | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |
| 75 years and over | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |

Figure 2: Mean Earnings of Asian Male



Standard Error of Mean

| Characteristic | Total | Less Than 9th Grade | High school | | College | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Some college No degree | Associate degree | Bachelor's degree or more | | | | |
| | | | 9th to 12th nongrad | Graduate (incl GED) | | | Total | Bachelor's degree | Master's degree | Professional degree | Doctorate degree |
| Total | 2,235 | (B) | (B) | 3,065 | 9,877 | 3,797 | 2,765 | 3,211 | 4,324 | 21,743 | 7,372 |
| Under 65 years | 2,278 | (B) | (B) | 2,717 | 10,333 | 3,896 | 2,812 | 3,273 | 4,400 | 23,575 | 7,624 |
| 18 to 24 years | 5,288 | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |
| 25 to 34 years | 4,305 | (B) | (B) | 4,737 | (B) | (B) | 5,123 | 4,779 | 8,067 | (B) | (B) |
| 25 to 29 years | 5,626 | (B) | (B) | (B) | (B) | (B) | 6,713 | 6,880 | 11,233 | (B) | (B) |
| 30 to 34 years | 6,019 | (B) | (B) | (B) | (B) | (B) | 6,910 | 6,024 | 10,592 | (B) | (B) |
| 35 to 44 years | 4,262 | (B) | (B) | 3,912 | (B) | (B) | 4,390 | 5,665 | 8,333 | (B) | 10,627 |
| 35 to 39 years | 6,069 | (B) | (B) | (B) | (B) | (B) | 7,061 | 8,740 | 14,585 | (B) | (B) |
| 40 to 44 years | 5,911 | (B) | (B) | (B) | (B) | (B) | 4,736 | 6,074 | 7,773 | (B) | (B) |
| 45 to 54 years | 4,623 | (B) | (B) | 3,538 | (B) | (B) | 5,841 | 5,560 | 7,449 | (B) | (B) |
| 45 to 49 years | 6,849 | (B) | (B) | (B) | (B) | (B) | 8,614 | 8,613 | 10,477 | (B) | (B) |
| 50 to 54 years | 5,544 | (B) | (B) | (B) | (B) | (B) | 6,651 | 6,585 | 8,369 | (B) | (B) |
| 55 to 64 years | 5,252 | (B) | (B) | 10,346 | (B) | (B) | 7,954 | 12,965 | 8,955 | (B) | (B) |
| 55 to 59 years | 5,820 | (B) | (B) | (B) | (B) | (B) | 7,947 | 7,078 | (B) | (B) | (B) |
| 60 to 64 years | 10,033 | (B) | (B) | (B) | (B) | (B) | 15,756 | 25,702 | (B) | (B) | (B) |
| 65 years and over | 13,085 | (B) | (B) | (B) | (B) | (B) | 15,752 | (B) | (B) | (B) | (B) |
| 65 to 74 years | 13,619 | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |
| 65 to 69 years | 13,489 | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |
| 70 to 74 years | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |
| 75 years and over | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) | (B) |

Figure 3: Standard Error of Asian Male

Data was pulled in the same way for men and women of each of the aforementioned races. Note that some categories have no mean earnings, because there was too small a sample size.

# Methodology

In this project, only workers who worked full-time and year-round were considered. Because the census data was provided as a mean value per category, the independent variables were best represented by dummy variables to produce a multilinear regression model. A weighted least squares method was implemented because each category's mean was created with a differing sample size. In addition to the final multilinear model, other R models and plotting software were used to further analyze the census data.

The following predictor variables were used for analysis:

- $X_1$: The categories of age
- $X_2$: Gender
- $X_3$: Education levels
- $X_4$: Race
- $Y$: Mean earnings

Dummy variables were used to indicate the predictor variables, each of which were k-level quantitative variables. Dummy variables, a set of binary indicators, represented each of the independent variables. Since there were 7 age categories, 6 dummy variables (or "dv's") were needed. Similarly, gender required 1 "dv," race required 3 "dv's," and education required 2 "dv's." The R function 'as.factor()' was used to implement each of these sets of variables. For example, race have 4-level, then it required 3 dummy variable equations:

$$X_{41} = \begin{cases} 1 & \text{, if person is Black} \\ 0 & \text{, otherwise} \end{cases}$$

$$X_{42} = \begin{cases} 1 & \text{, if person is Hispanic} \\ 0 & \text{, otherwise} \end{cases}$$

$$X_{43} = \begin{cases} 1 & \text{, if person is White} \\ 0 & \text{, otherwise} \end{cases}$$

Thus, there are total 12 predictor variables in the model where each $X_{ij}$ represents the dummy variable with associated weighting

$$X_{ij} = \begin{cases} 1 & \text{, if "belong in a level"} \\ 0 & \text{, otherwise} \end{cases}$$

folloing as the below table:

| Age($X_1$) | Gender($X_2$) | Education level ($X_3$) | Race ($X_4$) |
|---|---|---|---|
| $X_{11}$: 25-34 | $X_{21}$: Male | $X_{31}$: Bachelor's and more | $X_{41}$: Black |
| $X_{12}$: 35-44 | | $X_{32}$: High school | $X_{42}$: Hispanic |
| $X_{13}$: 45-54 | | | $X_{43}$: White |

| Age($X_1$) | Gender($X_2$) | Education level ($X_3$) | Race ($X_4$) |
|---|---|---|---|
| $X_{14}$: 55-64<br>$X_{15}$: 65-74<br>$X_{16}$: 75&up | | | |

Finally, the model was weighted using the standard error of mean. For each category's mean earning, the census provided the standard error of mean. The full R model with beta coefficient values can be found in appendix C.

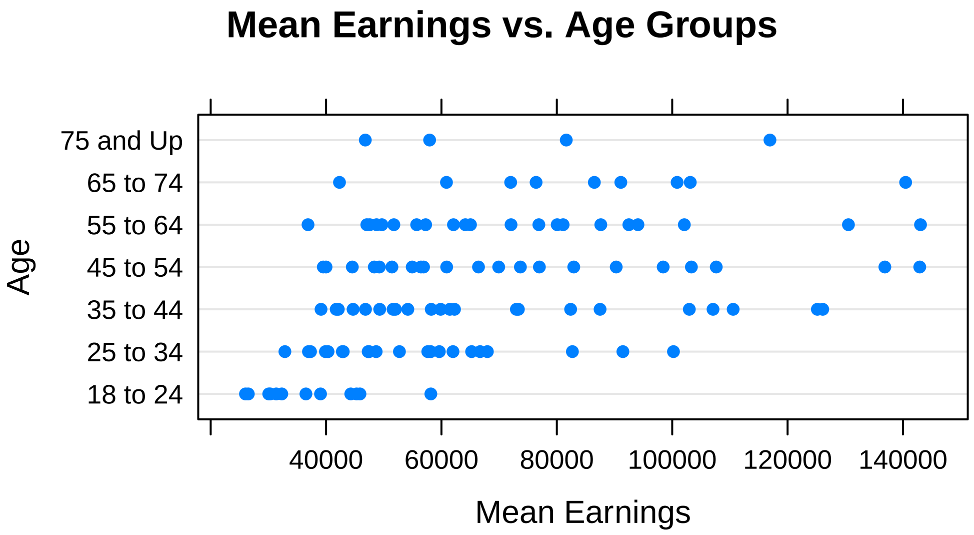
MLR model: $Y = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{21} + \beta_8 X_{31} + \beta_9 X_{32} + \beta_{10} X_{41} + \beta_{11} X_{42} + \beta_{12} X_{43} + \epsilon$

The model was then analyzed primarily with R visualization software, such as normal plots and Q-Q plots. Additionally, each independent variable was further analyzed by using SLR models and various R visualization tools. Finally, to better understand the relationships between the independent variables, further sources were utilized to show that correlation may not equate to causation. These additional sources are presented and discussed under the "Analysis and Results" section.

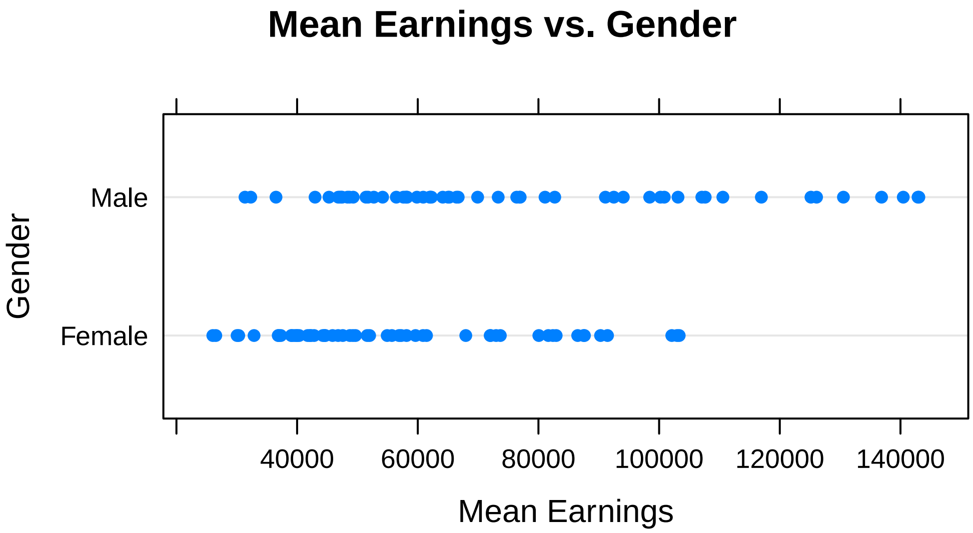
## Data Analysis and Results

1. Basic Visualization of Data

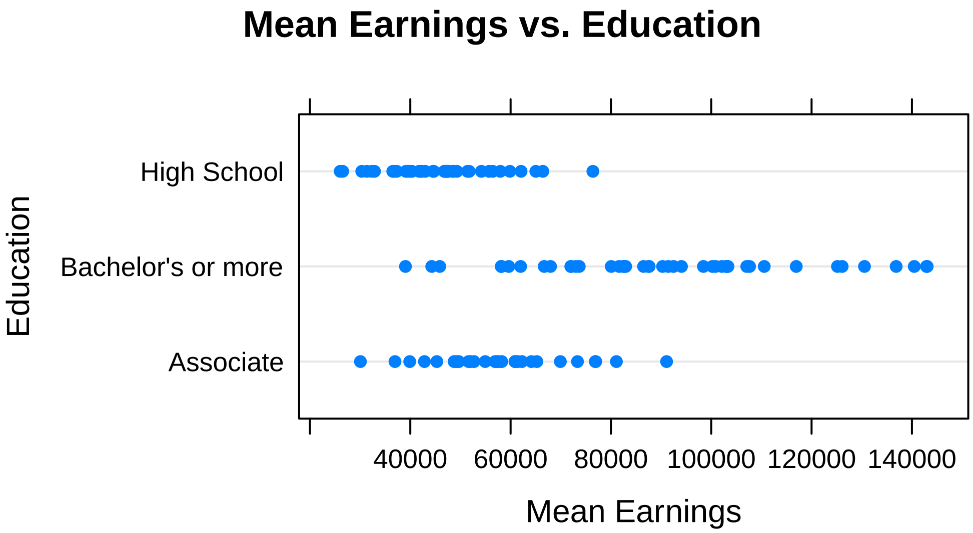
Prior to creating the weighted single linear regression and multilinear regression models, dot plots were created to have a very basic visual representation of the raw data. In the below dot plots, each dot is one group which falls in that category.



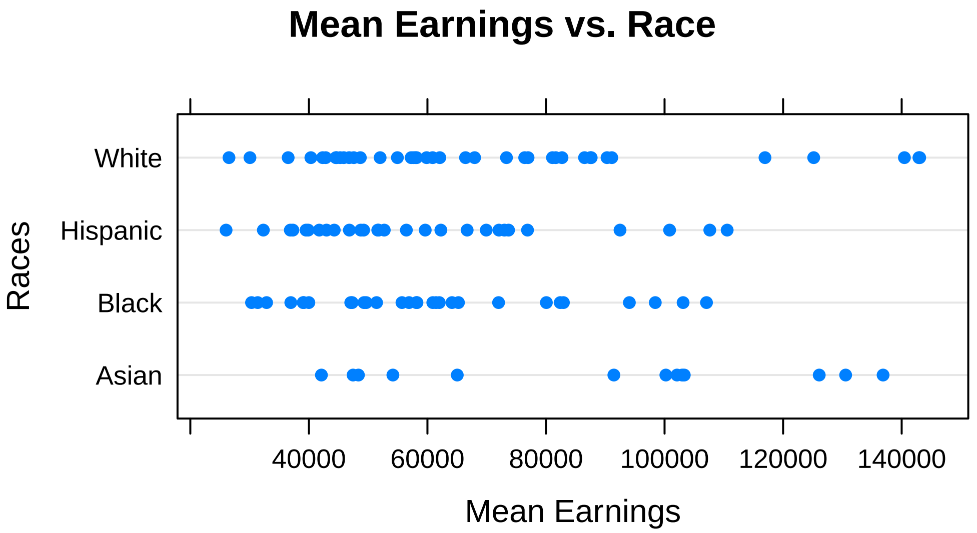
*Plot I: Mean Earnings vs. Age Groups.*



*Plot II: Mean Earnings vs. Gender.*



*Plot III: Mean Earnings vs. Education.*



*Plot IV: Mean Earnings vs. Race.*

It should be noted that, due to insufficient sample size in some groups, there is not an equal distribution of dots (groups) across all categories. For example, in Plot I the age "75 and up" category only has dots representing the following groups: {Male, High School, White}, {Male, Bachelor's or more, White}, {Female, High School, White}, and {Female, Bachelors or more, White}. Another important note is that each dot (group) has a different sample size, which may alter the perception of the range of mean earnings in each of the plots.

At a glance, several conclusions can be drawn. For the Plot I, it appears that mean salary has a much greater spread in older age groups compared to younger groups. Young people are concentrated in a low range, which is reasonable as they are early in their careers. From Plot II there appears to be a mean salary advantage for males. From plot III there is a limited mean salary range for those with only a high school level education where the highest group earns about 75k. For those with a Bachelor degree or higher, there's a broad mean salary range from about 40k-150k. Plot III seems to show that education level has the strongest relationship with expected mean earnings since it has the most distinctive groupings per category. Finally, from Plot IV there appears to be similar mean salary ranges for Hispanic and Black categories, and somewhat similar mean salary ranges for White and Asian categories.

Overall from looking at these plots, the greatest ranges for mean earnings are for White race and male gender. Thus (depending on the standard error for each group) there may be a poor statistical significance for those variables. There will likely also be poor statistical significance for variables with few samples, such as age group "75 and up."

Finally, it was noticed that not all "independent" variables were actually independent of each other, as shown below in Table 2.

Table 2: Independence of Gender and Education Level [2]

| Gender | Female | | Male | |
|---|---|---|---|---|
| Education Level | n(1000's) | Distribution | n(1000's) | Distribution |

| Gender | Female | | Male | |
|---|---|---|---|---|
| High School | 10662 | 26% | 71985 | 25% |
| Associate | 5933 | 14% | 84501 | 29% |
| Bachelor’s or more | 25125 | 60% | 131398 | 46% |

For gender and education level to be independent variables from each other, it would be expected that the same distributions of females and males are present across levels of education. The same was also found for race and education. The lack of full independence will likely lead to a skew in interpretation of the impact gender, race, or education levels may have on mean salary earnings.

2. Simple Linear Regression Analysis

Included in the analysis are the single linear regression models for both gender vs mean salary, and education level vs mean salary. Respectively, those were the weakest and strongest relationships between independent and dependent variables. The below analysis will include why the “Education Model” is more statistically significant than the “Gender Model.”

- “Education Model” - the simple regression test the effects of mean earnings on education levels. The Education Model uses the associate level variable as the dummy variable. The simple linear regression (SLR) equation: $Y_{\text{edu}} = \beta_{0'} + \beta_{1'}X_{31} + \beta_{2'}X_{32} + \epsilon$

After running SLR in R, we obtained the model: $\hat{Y}_{\text{edu}} = 58148 + 33259X_{31} - 12146X_{32}$, p-value $<$ 2.2e-16, and $R^2 = 0.5293$.

- “Gender Model” - the simple regression test the effects of mean earnings on gender. The Gender Model uses the female variable as the dummy variable (represented as the intercept). SLR equation: $Y_{\text{gender}} = \beta_{0''} + \beta_{1''}X_{21} + \epsilon$

After running SLR in R, we obtained the model: $Y_{\text{gender}} = 57245 + 20180X_{21}$, p-value $= 0.000166$, and $R^2 = 0.1225$.

The above SLR models use dummy variables. The Gender Model uses the female variable as the dummy variable (represented as the intercept), and the Education Model uses the associate level variable as the dummy variable. The Education Model appears to have a stronger reliability than the Gender model, since the P-value is smaller and the adjusted R-squared value is significantly larger. This would lead to reject the null hypothesis and conclude that there is strong evidence that a relationship between mean earnings and education.

In figure 4, for the Gender Model, females are the left grouping and males are on the right grouping. For the Education Model, from left to right is the high school, associates, and bachelors+ grouping. The Education Model seems to have a slightly better fit because the fitted residuals line is slightly flatter, indicating a distribution of residuals which more closely resembles the assumption of normality.

In Figure 5, the Normal Q-Q plot has a “S” shape and not following with residuals line; this indicates a lack of normality in the error. Overall, the residuals doesn’t look to have a normal distribution.

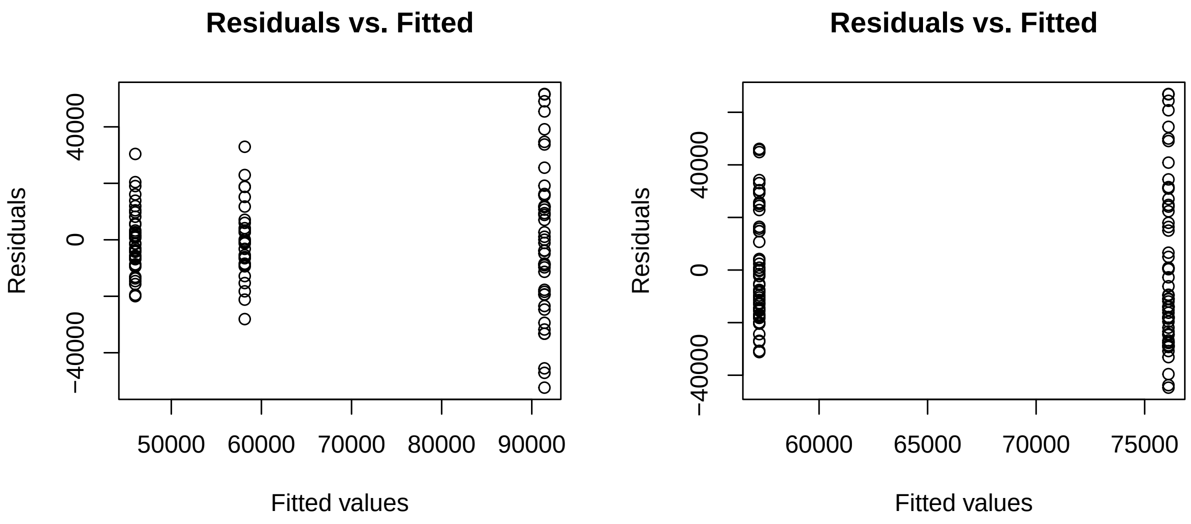


Figure 4: Residuals vs. Fitted of Education Model and Gender Model

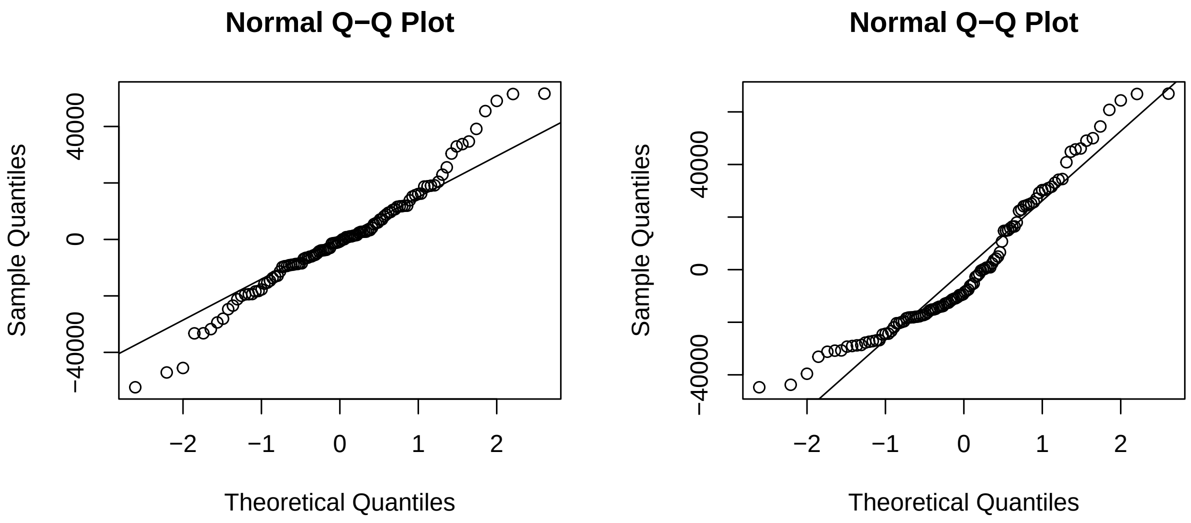
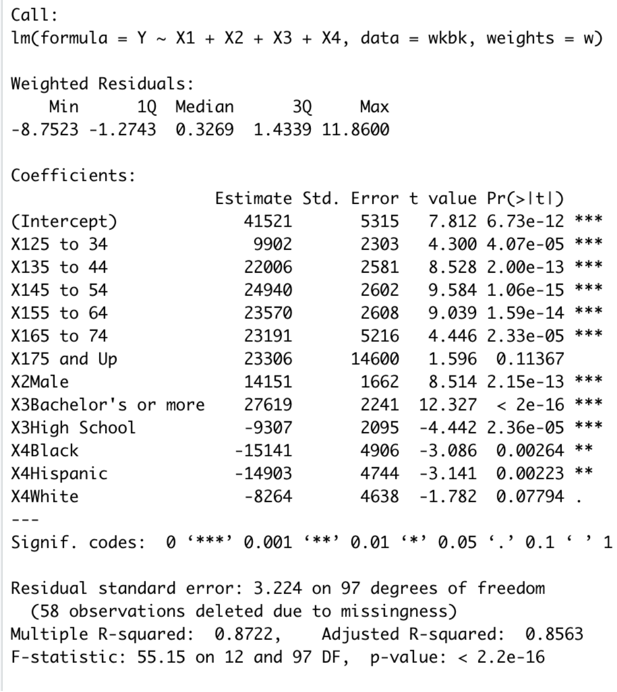


Figure 5: Normal Q-Q of Education Model and Gender Model

3. Multiple Linear Regression

As we mentioned in Methodology section, our model was implemented by using weighted least square method. Let $\sigma$ be standard errors obtained from the data, we can calculus weights as $w = \frac{1}{\sigma^2}$. Using the R function ‘*fit.weighted <- lm(Y ~ X1+X2+X3+X4,data=wkbk,weights = w)*’.



```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = wkbk, weights = w)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-8.7523 -1.2743  0.3269  1.4339 11.8600 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)             41521       5315   7.812 6.73e-12 ***
X125 to 34               9902       2303   4.300 4.07e-05 ***
X135 to 44              22006       2581   8.528 2.00e-13 ***
X145 to 54              24940       2602   9.584 1.06e-15 ***
X155 to 64              23570       2608   9.039 1.59e-14 ***
X165 to 74              23191       5216   4.446 2.33e-05 ***
X175 and Up             23306      14600   1.596  0.11367    
X2Male                  14151       1662   8.514 2.15e-13 ***
X3Bachelor's or more    27619       2241  12.327  < 2e-16 ***
X3High School           -9307       2095  -4.442 2.36e-05 ***
X4Black                -15141       4906  -3.086  0.00264 ** 
X4Hispanic             -14903       4744  -3.141  0.00223 ** 
X4White                 -8264       4638  -1.782  0.07794 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.224 on 97 degrees of freedom
  (58 observations deleted due to missingness)
Multiple R-squared:  0.8722,    Adjusted R-squared:  0.8563 
F-statistic: 55.15 on 12 and 97 DF,  p-value: < 2.2e-16
```

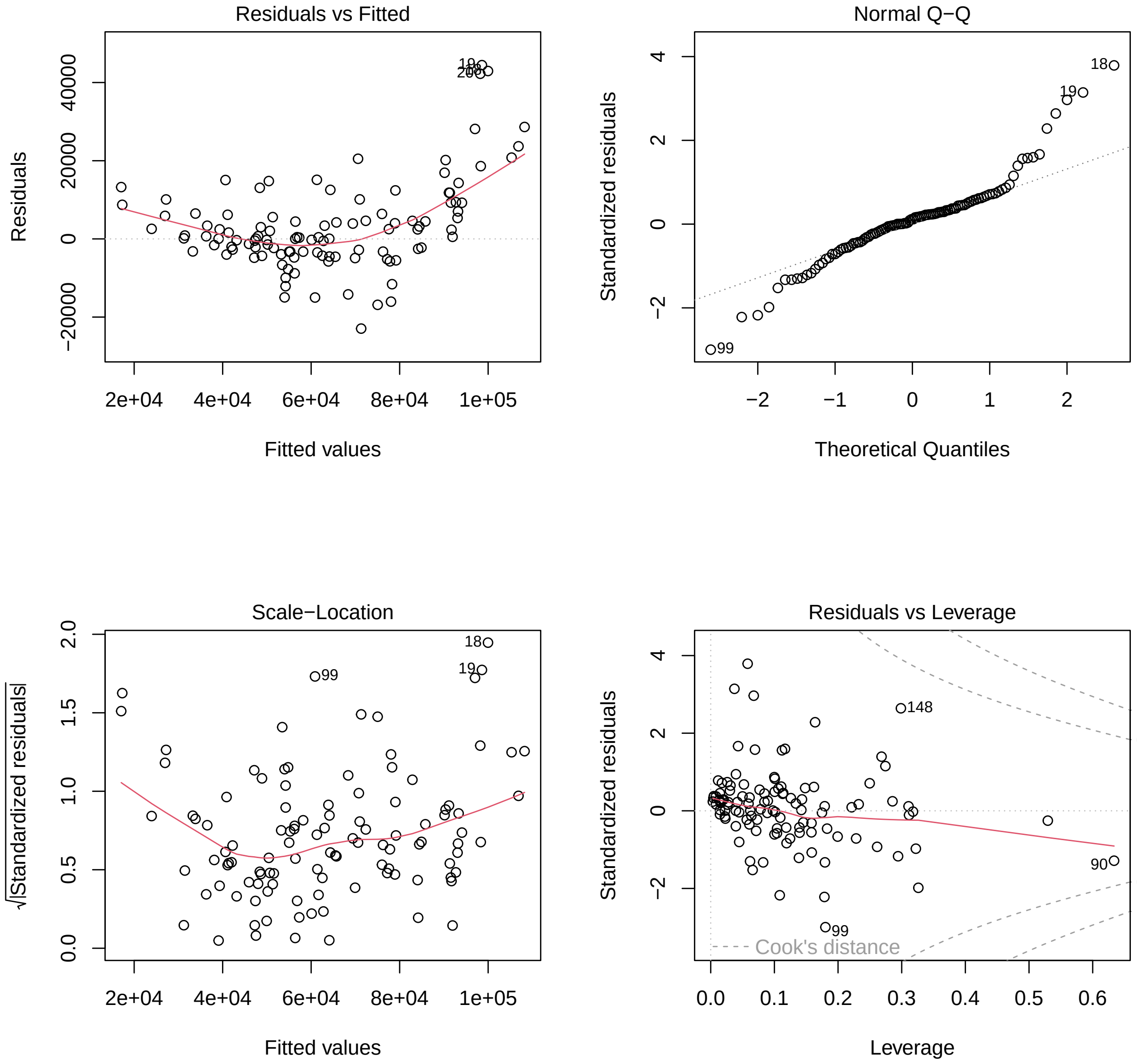
Figure 6: Summary of MLR model

We obtained the MLR model as below: $Y = 41521 + 9902X_{11} + 22006X_{12} + 24940X_{13} + 23570X_{14} + 23191X_{15} + 23306X_{16} + 14151X_{21} + 27619X_{31} - 9307X_{32} - 15141X_{41} - 14903X_{42} - 8264X_{43}$, p-value: $< 2.2\text{e-}16$, and $R^2 = 0.8722$.

From the MLR model created in R, the significance codes show lowest significance for the White group. This is to be expected from looking at Plot IV: Race vs Mean Earnings, since there is such a large range of mean salary earners amongst the White group. Also note that the standard error is highest for the age group ‘75 and up,’ which was also expected from Plot I: Age Groups vs Mean Earnings, since there are very few data points. From the large F-statistic and small p-value, it can be concluded that the MLR model is, overall, statistically significant.

Considering our ‘baseline’ for dummy variables are ‘age 18 to 24,’ female, associates level, and asian, the following explanation will say whether there is an expected increase or decrease in mean salary earnings when those variables change. Overall, there are significant and positive relationships between mean earnings and the following variables: all age groups except for age ‘75 and up,’ male, and bachelor’s or greater. There are significant and negative relationships between mean earnings and the following variables: high school level, Black group, and Hispanic group. No relationship can be concluded for age ‘75 and up’ or for White group.

A residual plot was created to better understand if the MLR model follows the normal assumptions, and if the linear fit is the best fit. We designed the diagnostic plots for our model individually, but there is a function in R that allows us to create all of these plots at once.



Due to the curvature of the Residuals vs Fitted plot, it can be concluded that there's a quadratic relationship between the independent variables and the mean income, meaning our linear assumption is not the best fit. Moving from left to right on the graph, there is a slightly greater spread of residuals which shows heteroskedasticity. Heteroskedasticity means that the variance of errors is not constant, which goes against one of the assumptions for our multilinear regression model.

The Normal Q-Q plot has a slight "S" shape, with residuals dipping below and above the dotted line; this indicates a lack of normality in the error. The multilinear regression model assumes a normal distribution of error, so this graph shows that the model is also slightly imperfect, though not completely unreliable since most of the residuals fall on the line.

The p-value = 2.2e-16, the null hypothesis is rejected, so we can conclude that our model is more effective than a model only with the intercept since at least one coefficient is significantly different from 0. Further, R-square = 0.8722, our model is fitting the data quite well.

It must be noted that correlations may not provide true evidence of causation. For example, mean income level typically varies around the US by region and by size of city. Racial makeup of the US is also not evenly distributed and will vary by region. The below tables show just a few example cities, demonstrating how demographics and mean salary vary.

Table 3: Racial Demographic by City [1]

| Region | Asian | Black | Hispanic | White | Population |
|---|---|---|---|---|---|
| USA Total | 6.1% | 13.6% | 18.9% | 59.3% | 331,449,281 |
| Atlanta, GA | 4.8% | 49.8% | 4.9% | 38.0% | 498,602 |
| New York, NY | 14.3% | 23.8% | 28.9% | 31.9% | 8,804,190 |
| Seattle, WA | 16.3% | 7.1% | 7.1% | 62.6% | 735,015 |
| Chicago, IL | 6.8% | 29.2% | 28.6% | 33.3% | 2,747,231 |
| Des Moines, IA | 6.5% | 11.4% | 14.0% | 64.6% | 214,137 |

Table 4: Salary by City [3]

| City | Weighted Mean | Weighted Median |
|---|---|---|
| Atlanta | $76,387.48 | $63,028.90 |
| New York | $85,387.63 | $80,955.84 |
| Chicago | $66,614.71 | $49,539.49 |
| Seattle | $75,010.09 | $74,631.17 |
| Des Moines | $55,084.19 | $50,556.51 |

For this case, the sample data may be more reliably interpreted if it was isolated within one region or city. To point out how the US-wide census may draw unreliable conclusions, note how both Chicago and New York have significantly different weighted mean salaries, yet have similar racial demographics. Meanwhile, Seattle and Atlanta have very similar mean salaries but vastly different racial demographics.

## Conclusions

The weighted MLR model was found to be statistically significant for 10 out of 12 dummy variables, with no significant relationship concluded for variables 'age 75 and up' or White group. The MLR model did not perfectly follow all MLR assumptions; residuals mostly followed a normal distribution though it was imperfect, it was noted that a quadratic relationship would likely provide a better fit than a linear model, and not all "independent" variables may have been independent from each other. Overall the model was determined to be statistically significant, but had room for improvement.

From the model, it's of course important to interpret the results for practical implementation. Does education level greatly impact salary expectations? Is there gender discrimination against certain genders or certain races? Unfortunately this project was not expansive enough to draw such conclusions. Racial demographics and mean salary vary significantly by region. Since population size of regions also vary, it may be easy to draw false conclusions from results. Even though racial and gender demographics show clear differences, there could also be further analysis for job type as well, perhaps more women are highly educated compared to men, but perhaps the education may fall in fields that result in lower demand or lower salaries. The lack of practical conclusions serve as a starting point for further research and statistical analysis to determine if discrimination is present. Variables would have to be further isolated to be better interpreted.

# Appendix

**Appendix A**: Summary SLR of “Education Model”

```
##
## Call:
## lm(formula = Y ~ X3, data = wkbk)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52368  -9332   -704  10256  51638
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             58148       3655  15.910  < 2e-16 ***
## X3Bachelor’s or more    33259       4696   7.082 1.57e-10 ***
## X3High School          -12146       4790  -2.536   0.0127 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 19340 on 107 degrees of freedom
##   (58 observations deleted due to missingness)
## Multiple R-squared:  0.5293, Adjusted R-squared:  0.5205
## F-statistic: 60.16 on 2 and 107 DF,  p-value: < 2.2e-16
```

**Appendix B**: Summary SLR of “Gender Model”

```
##
## Call:
## lm(formula = Y ~ X2, data = wkbk)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44733 -18142  -8946  17585  66948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    57245       3592  15.937  < 2e-16 ***
## X2Male         18852       5034   3.745 0.000292 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 26400 on 108 degrees of freedom
##   (58 observations deleted due to missingness)
## Multiple R-squared:  0.1149, Adjusted R-squared:  0.1067
## F-statistic: 14.02 on 1 and 108 DF,  p-value: 0.0002916
```

**Appendix C**: Multilinear model in R with associated coefficients.

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = wkbk, weights = w)
##
## Coefficients:
##          (Intercept)            X125 to 34            X135 to 44
##                41521                  9902                 22006
##           X145 to 54            X155 to 64            X165 to 74
##                24940                 23570                 23191
##          X175 and Up                X2Male  X3Bachelor’s or more
##                23306                 14151                 27619
##       X3High School               X4Black            X4Hispanic
##                -9307                -15141                -14903
##              X4White
##                -8264
```

**Appendix D**: Summary of the MLR

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = wkbk, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7523 -1.2743  0.3269  1.4339 11.8600
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             41521       5315   7.812 6.73e-12 ***
## X125 to 34               9902       2303   4.300 4.07e-05 ***
## X135 to 44              22006       2581   8.528 2.00e-13 ***
## X145 to 54              24940       2602   9.584 1.06e-15 ***
## X155 to 64              23570       2608   9.039 1.59e-14 ***
## X165 to 74              23191       5216   4.446 2.33e-05 ***
## X175 and Up             23306      14600   1.596  0.11367
## X2Male                  14151       1662   8.514 2.15e-13 ***
## X3Bachelor’s or more    27619       2241  12.327  < 2e-16 ***
## X3High School           -9307       2095  -4.442 2.36e-05 ***
## X4Black                -15141       4906  -3.086  0.00264 **
## X4Hispanic             -14903       4744  -3.141  0.00223 **
## X4White                 -8264       4638  -1.782  0.07794 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.224 on 97 degrees of freedom
##   (58 observations deleted due to missingness)
## Multiple R-squared:  0.8722, Adjusted R-squared:  0.8563
## F-statistic: 55.15 on 12 and 97 DF,  p-value: < 2.2e-16
```

**Appendix E**: ANOVA Table of the MLR

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## X1         6 2203.7  367.29  35.3300 < 2.2e-16 ***
## X2         1  268.6  268.63  25.8401 1.804e-06 ***
## X3         2 4165.0 2082.48 200.3151 < 2.2e-16 ***
## X4         3  242.6   80.86   7.7778 0.0001046 ***
## Residuals 97 1008.4   10.40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## References

[1] US Census Bureau. (2021, April). U.S. Census Bureau quickfacts: Chicago City, Illinois. QuickFacts. Retrieved November 2022, from https://www.census.gov/quickfacts/chicagocityillinois

[2] US Census Bureau. (2020, November 15). PINC-04. educational attainment–people 18 years old and over, by total money earnings, work experience, age, race, Hispanic origin, and sex. Census.gov. Retrieved November 2020, from https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-pinc/pinc-04.2020.html#list-tab-T5LKCZT8B1NNRPILCH

[3] Golden Oak Research Group. (2018, April 16). US household income statistics. Kaggle. Retrieved November 2022, from https://www.kaggle.com/datasets/goldenoakresearch/us-household-income-stats-geo-locations?resource=download