Potential Contributors to NYC Income

Nathan Yao nyao

Due Monday, July 15, at 11:59PM

Contents

Introduction

New York City. Nicknamed The Big Apple, this city consists of multiple burroughs that has a diverse group of people. The city is known to be one of the most influential cities in America due to a variety of factors such as the economic opportunities. However, an issue that has been present in recent years has been the concentrated population in New York City, causing housing to be a prominent issue many constituents are trying to solve today. In order for people to afford such housing prices, they have to have a certain annual income. In this report, we look at how an NYC household's income is influenced by three different factors, age of the respondent, the number of maintenance deficiencies on their house, and the year that the responded move to New York City.

Exploratory Data Analysis

Introduction to the Data: The data that is going to be analyzed is provided by The United States Census Bureau, and sponsored by the New York City Department of Housing Preservation and Development. The poll took into account of 4 variables with a total of 299 respondents in New York City. This paper will analyze the respondent's total household income as the response variable and 3 explanatory variables that consist of the respondent's age, number of maintenance deficiencies of the residence, between 2002 and 2005, and the year the respondent moved to New York City. The variables are described as:

Income: Total household income (in $) [the response variable] **Age:** Respondent's age (in years) **MaintenanceDef:** Number of maintenance deficiencies of the residence, between 2002 and 2005 **NYCMove:** The year the respondent moved to New York City

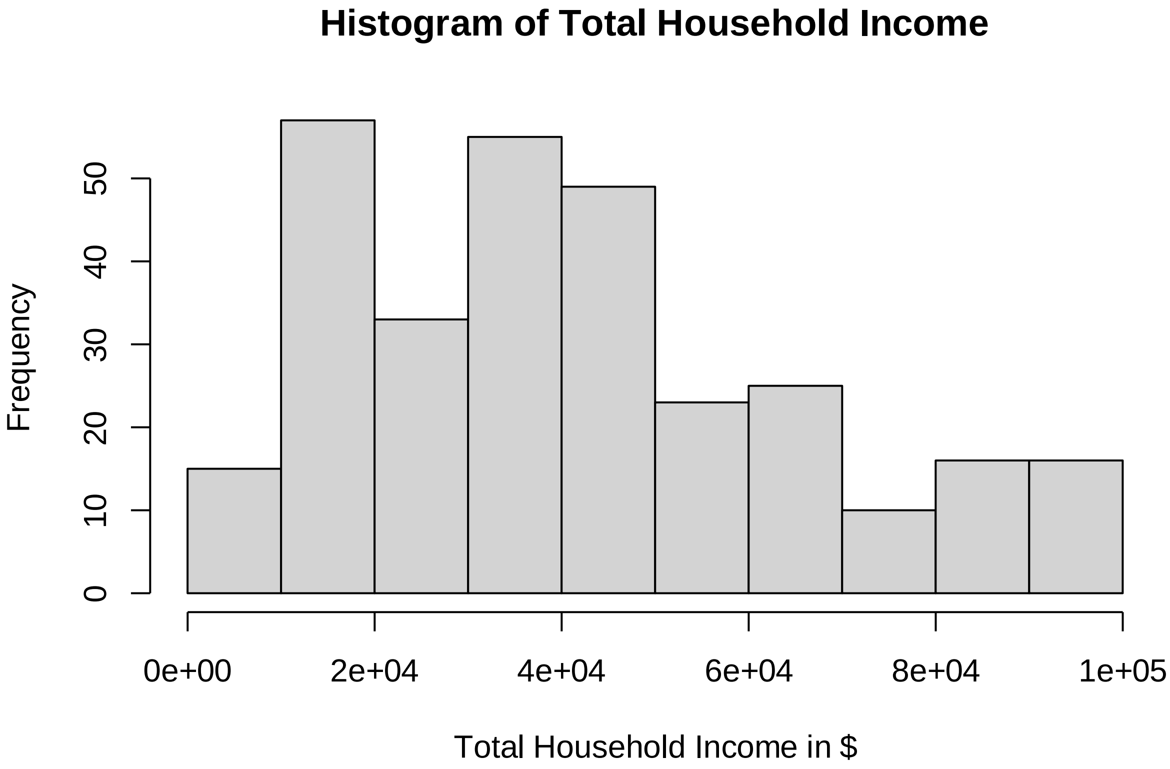
The data is presented in the fashion below (first 6 respondents):

```
head(nyc)
```

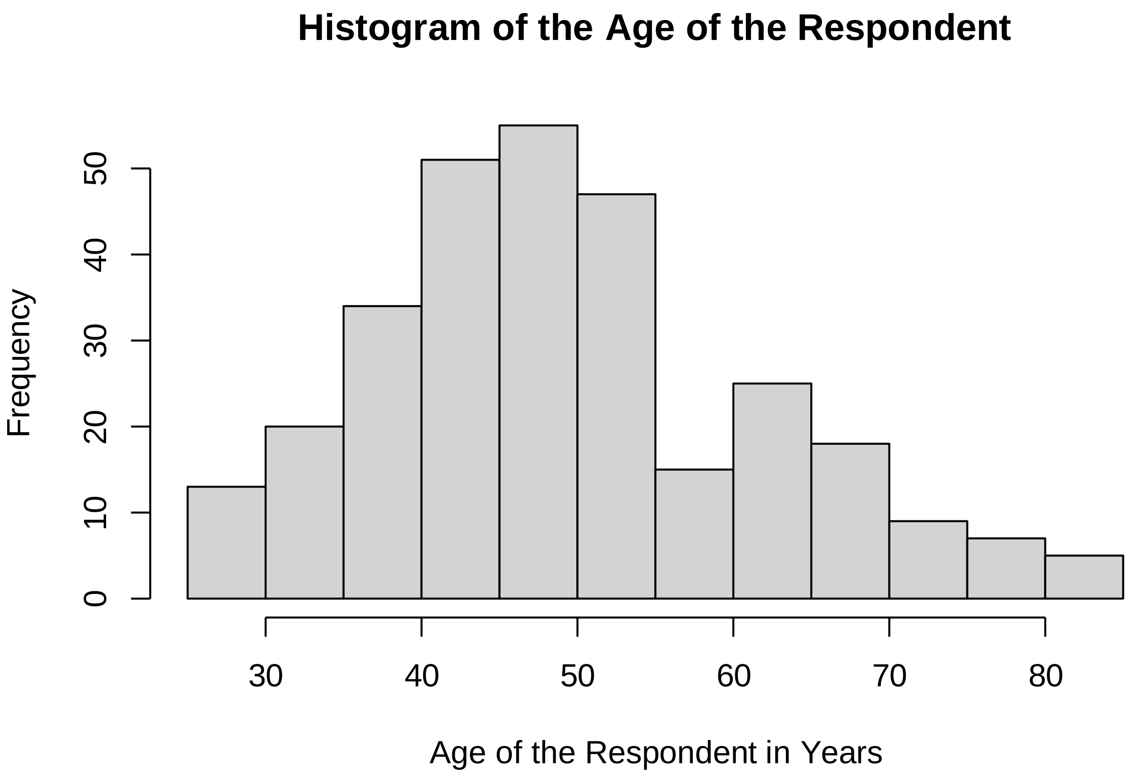
```
## # A tibble: 6 x 4
##    Income   Age MaintenanceDef NYCMove
##     <dbl> <dbl>          <dbl>   <dbl>
## 1    8400    77                   1      1981
## 2   17510    53                   2      1986
## 3   19200    33                   4      1992
## 4   42717    55                   1      1969
## 5    5000    58                   2      1989
## 6   30000    29                   4      1994
```

Univariate Analysis: Now we analyze each variable individually through univariate analysis. All of the variables are quantitative variables which means we will produce histograms of each variable's distribution along with summary data afterwards.

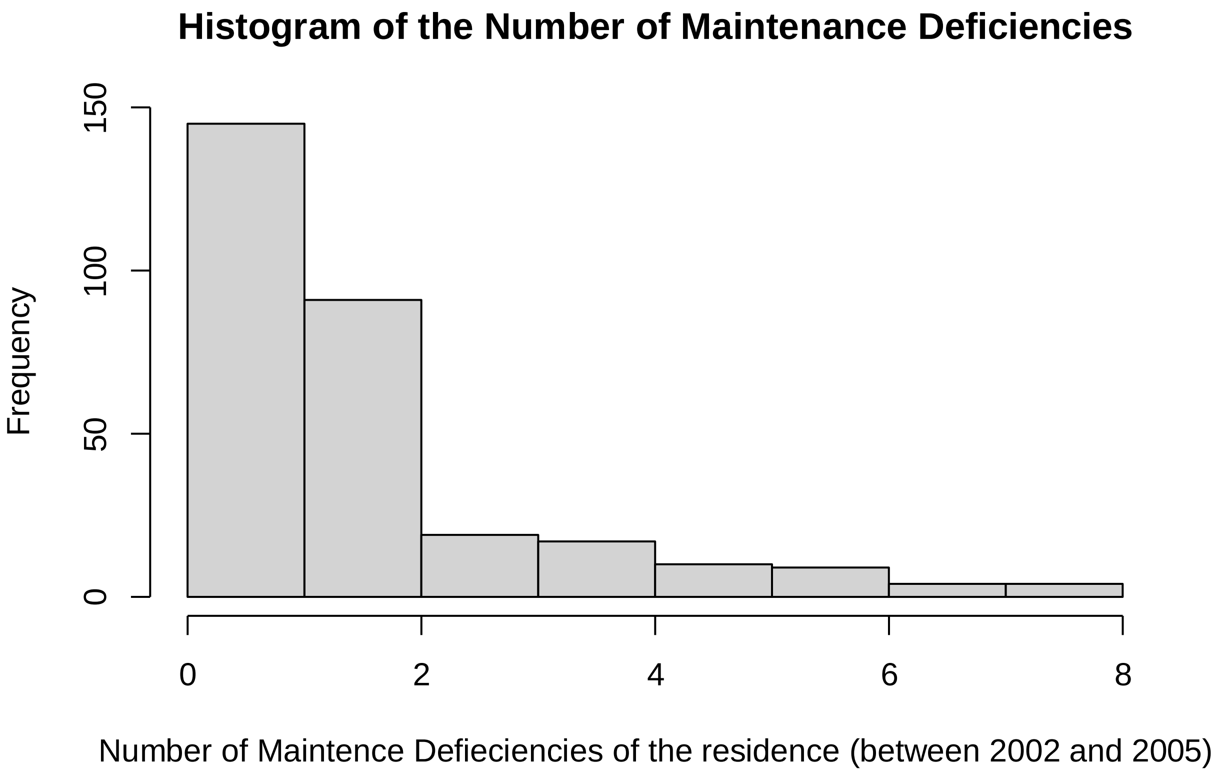
```
hist(nyc$Income, main = "Histogram of Total Household Income", xlab = "Total Household Income in $")
```



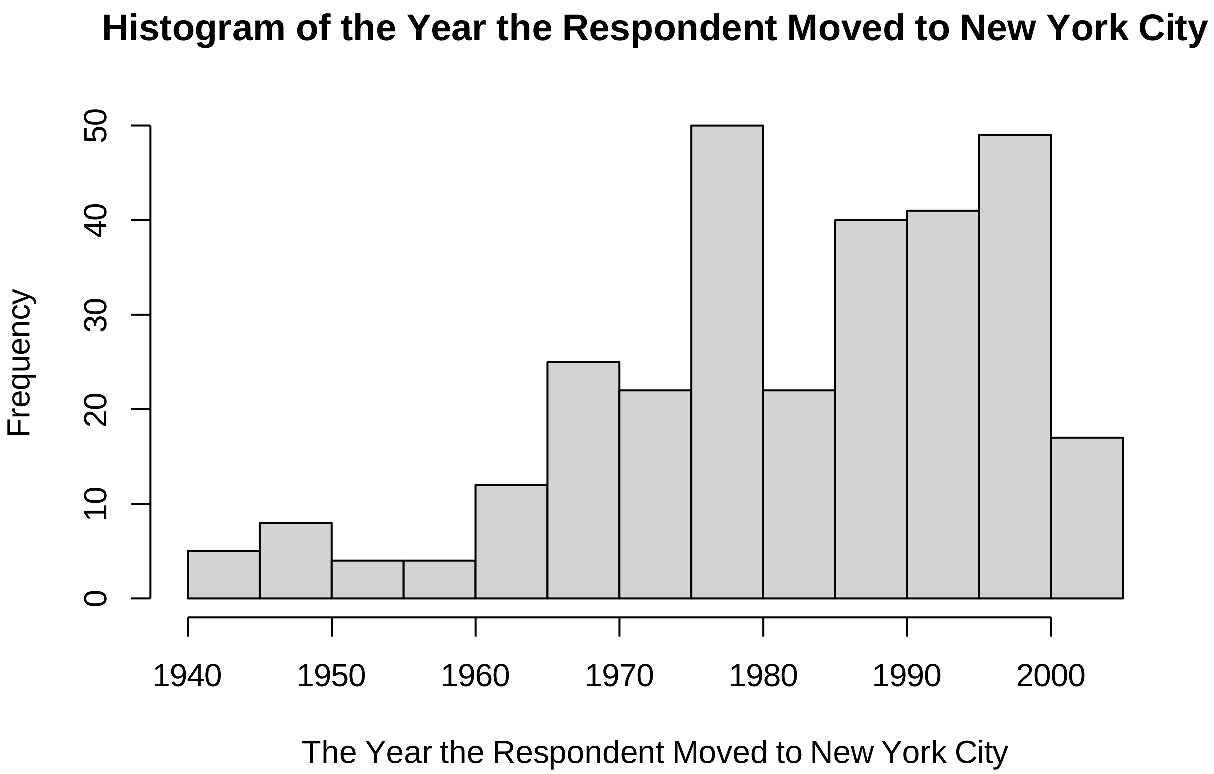
```
hist(nyc$Age, main = "Histogram of the Age of the Respondent", xlab = "Age of the Respondent in Years")
```



```
hist(nyc$MaintenanceDef, main = "Histogram of the Number of Maintenance Deficiencies", xlab = "Number o
```



```
hist(nyc$NYCMove, main = "Histogram of the Year the Respondent Moved to New York City", xlab = "The Yea
```



Here are the numerical summaries of the distributions above: For Income:

```
summary(nyc$Income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1440   21000   39000   42266   57800   98000 
```

For Age:

```
summary(nyc$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   26.00   42.00   49.00   50.03   58.00   85.00 
```

For Maintenance Deficiency:

```
summary(nyc$MaintenanceDef)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    1.00    2.00    1.98    2.00    8.00 
```

For the year that the respondent moved to New York City:

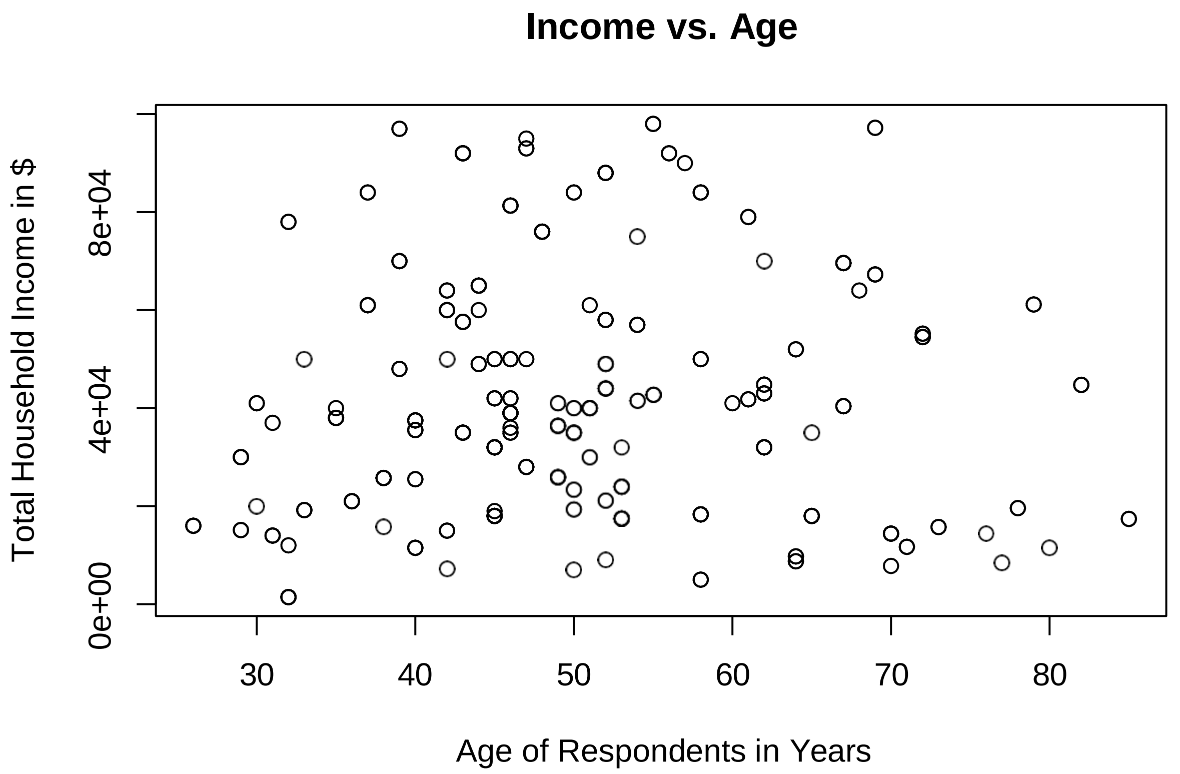
```
summary(nyc$NYCMove)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1942    1973    1985    1983    1995    2004 
```

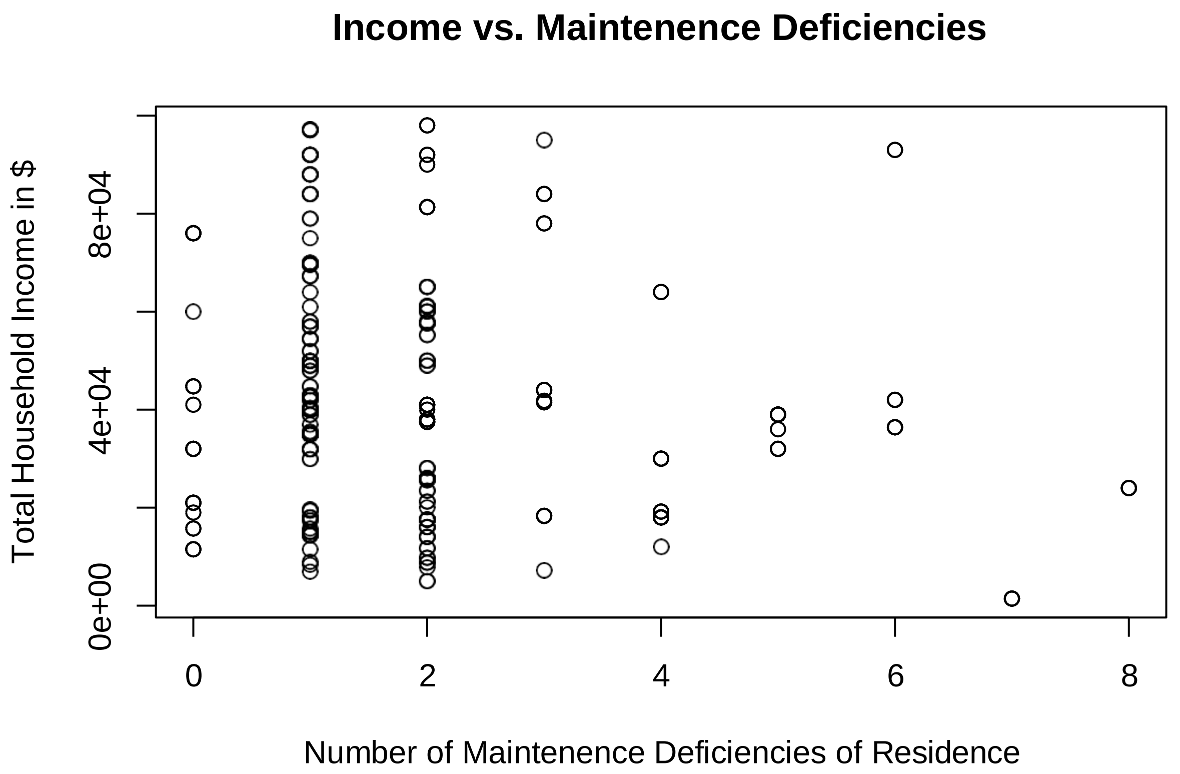
Here are the Observations of analyzing both the graphs and the summary statistics of each variable individually: The distribution of **total household income** roughly unimodal and right skewed. The Mean is just slightly larger than the median because of the right skew and there does not seem to be any outliers. The distribution of the age of the respondent is unimodal and rougly symmetric with a very slight right skew. The range of the age of the respondents ranges from 26 to 85 years old with a median and mean around 49 and 50 years old respectively. Furthermore, the distribution of the maintenance deficiencies per household is unimodal with a strong right skew. The range of the number of deficiencies ranged all the way from 0 deficiencies all the way to 8 deficiencies with a standard deviation of 1.62 deficiencies. If we look at the distribution of the years that respondents moved to New York City, we can see that the distribution is bimodal with peaks at 1970-1980s and after the 90s. The distribution is also left-skewed with a mean and median of 1983 and 1985 respectively.

Bivariate Exploration: After looking at the individual distribution of the variables, the next step is to look at possible relationships between income and the other 3 explanatory variables. The following are 3 scatterplots that demonstrate the relationships:

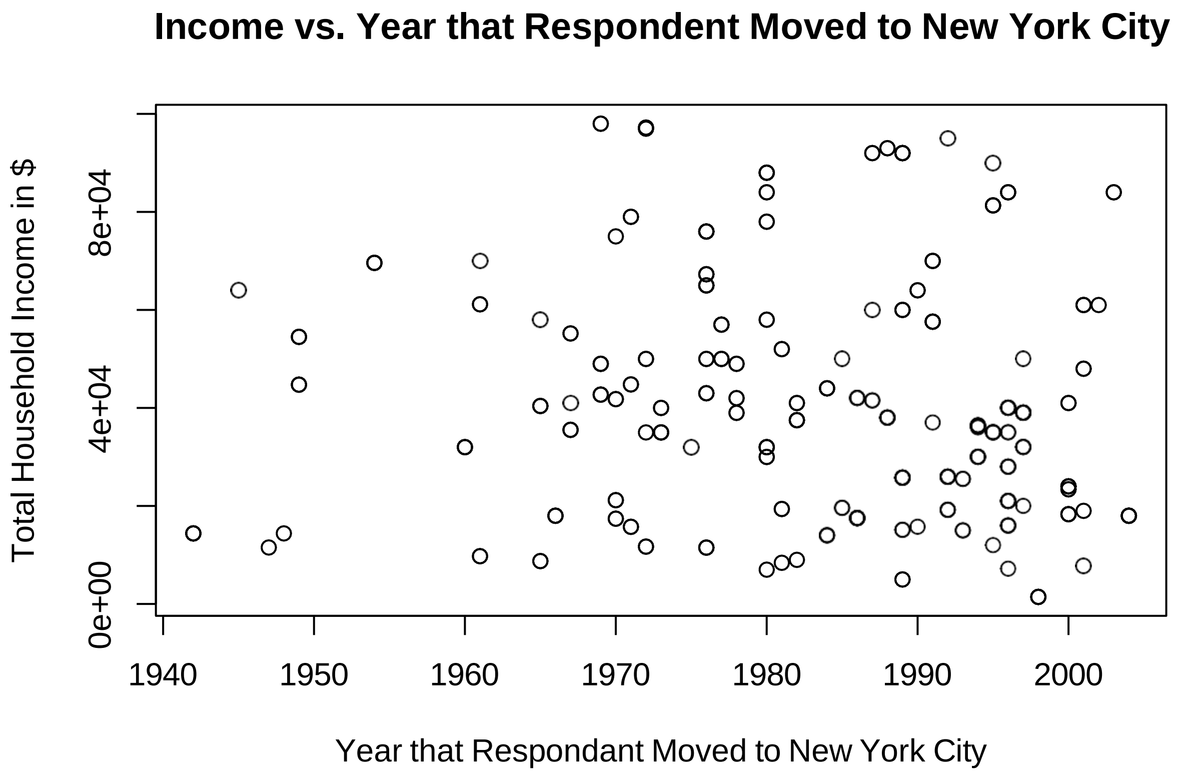
```
plot(Income~Age, data = nyc, main = "Income vs. Age", xlab = "Age of Respondents in Years", ylab = "Tot
```



```
plot(Income~MaintenanceDef, data = nyc, main = "Income vs. Maintenence Deficiencies", xlab = "Number of
```



```
plot(Income~NYCMove, data = nyc, main = "Income vs. Year that Respondent Moved to New York City", xlab
```

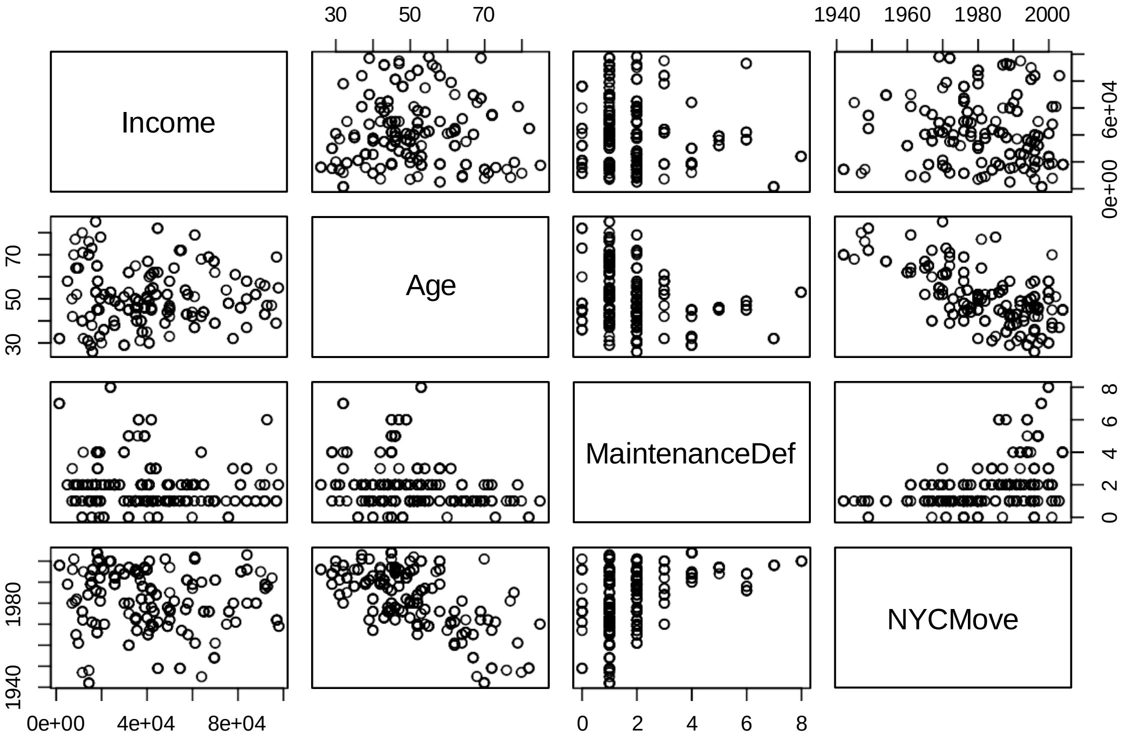


From the graphs above, the relationships do not look that strong. In the Income vs. Age scatter plot there does not seem to be a sign of any relationship between the two variables. The data points are concentrated near the 40-50 years mark. In the scatter plot with the number of Maintenance Deficiencies, we can see that there are a lot of data points at the 1 and 2 deficiency mark. There does seem to be a weak negative relationship between the two variables in this scatter plot compared to the other two scatter plots. Lastly, there also does not seem to be a strong relationship between the year that the respondent moved to New York City and the total household income. The distribution is left skewed with most of the data points after the 1970s or so.

Modeling

After analyzing the relationships between the variables, we now build a linear regression model that confirms some of the observation made above. More specifically, there does not seem to be a relationship between the total household income and the three explanatory variables. Before building the linear model, analyzing and completing the requirements for such a test is vital. We first explore if there is multicollinearity between the 3 explanatory variables, age of the respondent, the number of maintenance deficiencies per household, and the year that the respondent moved to New York City. Below is the pairs plot between the different variables:

```
pairs(nyc)
```



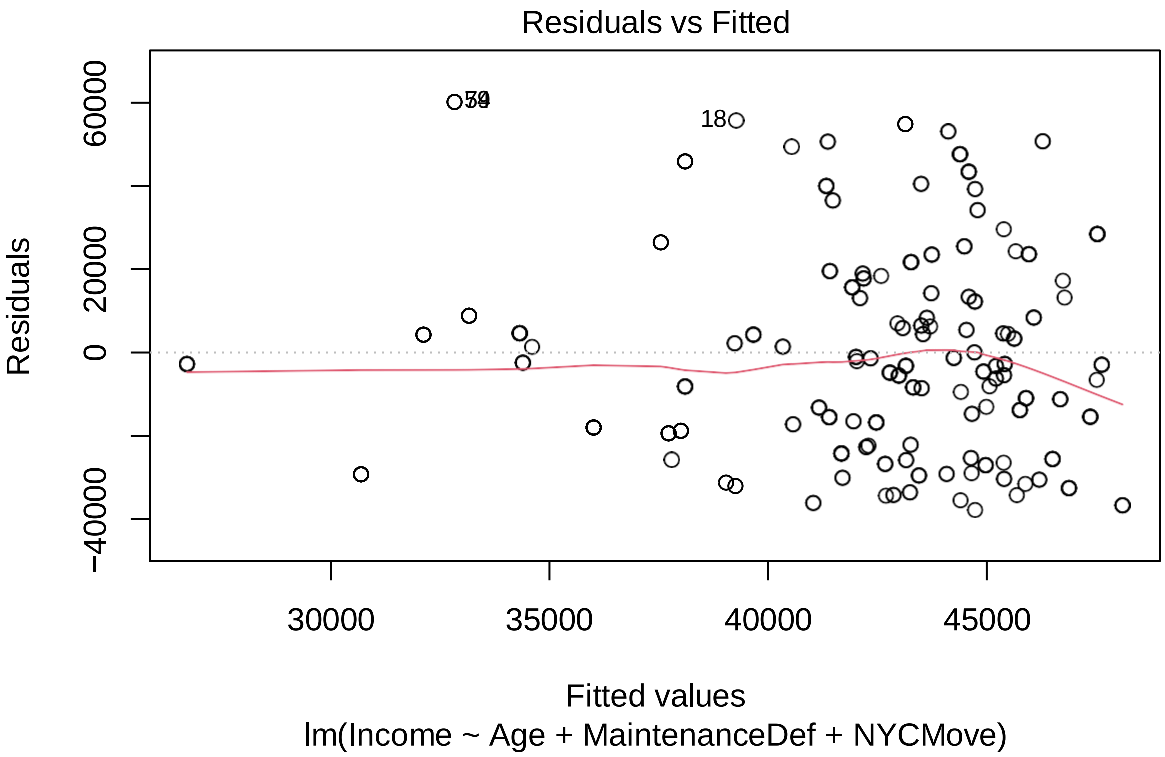
From the pairs plot above, we can observe that there might be possible multicollinearity issues. More specifically, when comparing Age and the number of maintenance deficiencies, there is a moderate negative relationship. Furthermore, and more importantly, Age and the year the respondent moved to New York City also has a strong negative relationship that may concern us regarding multicollinearity. To confirm this, we formally look at variation inflation values (VIF):

```
nyc.mod1 <- lm(Income~Age + MaintenanceDef + NYCMove, data = nyc)
car::vif(nyc.mod1)
```

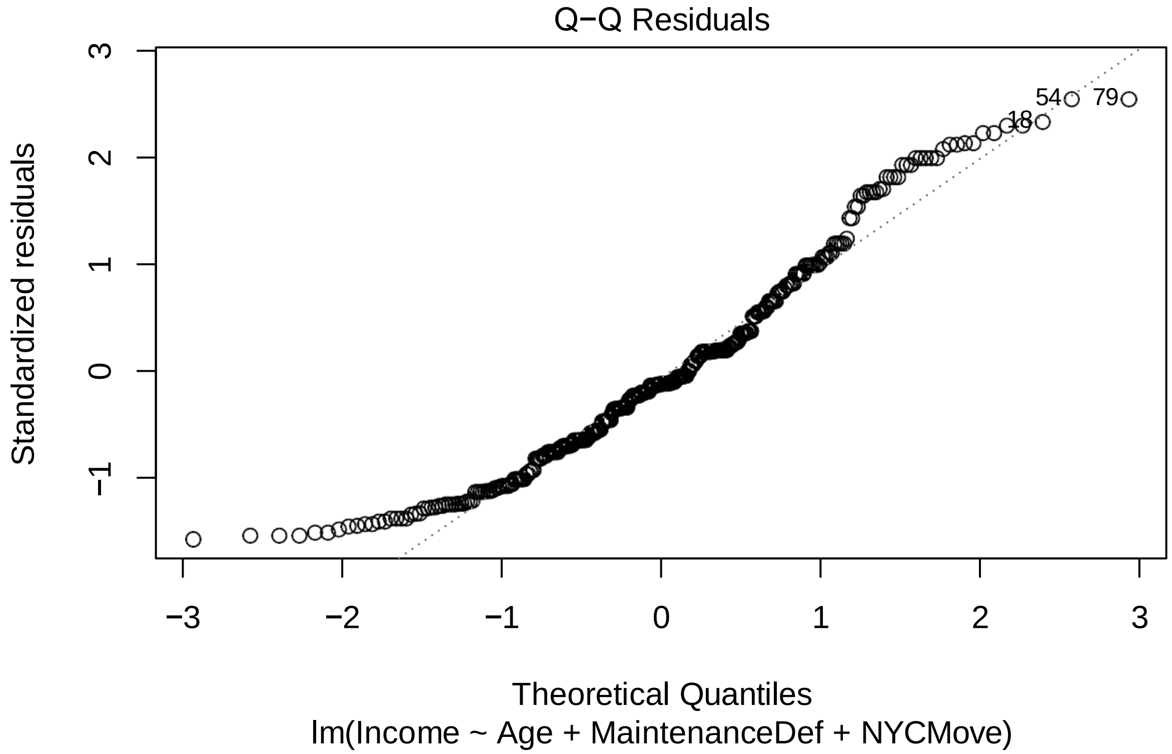
```
##            Age MaintenanceDef        NYCMove
##       1.687649       1.267728       1.999724
```

As we can see from the VIF values, none of the values are over 2.5 which means that the multicollinearity between the variables is not too problematic. We can continue to check our other requirements for the linear regression model. Below is the residual plot along with the normal qq-plot to analyze error requirements for the model:

```
plot(nyc.mod1, which = 1)
```



```
plot(nyc.mod1, which = 2)
```



From the residuals vs. fitted plot, we note that the residuals are fairly random without any pattern in the distribution above which satisfies independence between the residuals. Furthermore, the mean of the residuals are roughly 0 and a pretty constant standard deviation across the plot except for the clutter of points above and below near the 45000 mark. Moving on the normal qq-plot, it looks like there is a slight deviation off the line at the ends of the plot but nothing too problematic and we can move on with the linear regression model. Here is the regression analysis summary:

```
summary(nyc.mod1)
```

```
##
## Call:
## lm(formula = Income ~ Age + MaintenanceDef + NYCMove, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37734 -18010  -2878  14971  60171
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    237408.41  278939.01   0.851   0.3954
## Age               -71.98     144.97  -0.496   0.6199
## MaintenanceDef  -2273.22     964.72  -2.356   0.0191 *
## NYCMove           -94.34     138.82  -0.680   0.4973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23960 on 295 degrees of freedom
## Multiple R-squared:  0.02981,    Adjusted R-squared: 0.01995
## F-statistic: 3.022 on 3 and 295 DF,  p-value: 0.03005
```

This is the best linear regression model using the data given for a number of reasons. Although we could have made some transformations to the data, I decided not to because sometimes the model would be unstable. Furthermore, other models were tested and they included a variety of linear combinations of age, number of maintenance deficiencies, and the year the respondant moved to New York City. From comparing the models, the model with the highest R^2 value was still the model from above even though the R^2 value is only 0.02981.

The model is significant from the F-statistic p-value which is equal to $0.03005 < 0.05 = a$. However, only the p-value for the number of maintenance deficiencies is the only significant p-value out of the explanatory variables.

This confirms our observations in the bivariate analysis. Although all of the relationships of the explanatory variables and the total household income were somewhat weak, the p-value controlling for age and year that the respondent moved to New York City is statistically significant (p-value = 0.0191). This confirms that the scatterplot of the number of maintenance deficiencies predicting total household income as the most obvious relationship out of the explanatory variables.

Looking at the beta values all of the explanatory variables, the slopes are all negative, and looking at the bivariate scatter plots, the slopes being negative matches to some extent.

The model does not have any multicollinearity issues and matches the scatter plots in the bivariate analysis even thought the scatter plot relationships were not very obvious. The model also does not have a large R^2 value, only accounting for 2.981% of the variation of total household income. However, from the f-statistic p-value, total household income is associated with the age of the respondent, the number of maintenance deficiencies per household, and the year that the respondent moved to New York City.

Prediction

After establishing a model, we can look at the clients interest in predicting the income for a household with three maintenance deficiencies, whose respondent's age is 53 and who moved to NYC in 1987. To predict the value using our model above, we look at the beta/intercept values to create an equation. The calculation of such respondent is as follows:

```
237408.41-71.98*53-2273.22*3-94.34*1987
```

```
## [1] 39320.23
```

We predict that the income for a household with three maintenance deficiencies, whose repondent's age is 53, and who moved to NYC in 1987 is 39320.23 dollars per year.

Discussion

After completing this report, we can make many observations on the result and the data that was presented for the model. The data in the univariate and bivariate analysis did not seem to have any strong or obvious relationships. However, from the linear model above, there is actually some association between total household income and the three explanatory variables. A main concern and limitation that this model has is the R^2 value being very low. However, this just means that for the future, it is needed to find other variables that better predict total household income.