Predicting Income of NYC Housing Residents from Age, Maintainence Deficiencies, and Year of Move

Suha Niyas
sniyas

Due Wed, March 24, at 8:00PM (Pittburgh time)

Contents

```r
library("knitr")
library("cmu202")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("jtools")
library("leaps")
```

```r
social <- readr::read_csv("http://stat.cmu.edu/~gordonw/social.csv")
bikes <- readr::read_csv("http://stat.cmu.edu/~gordonw/bikes.csv")
nyc <- readr::read_csv("http://stat.cmu.edu/~gordonw/nyc.csv")
court <- readr::read_csv("http://stat.cmu.edu/~gordonw/court.csv")
```

Introduction

Despite being the largest city in the United States, and one of the most prominent places in the world, New York City has its flaws, especially in the residence department. Residents have to pay an exorbitant amount of money to rent out a one-bedroom apartment. Every three years, the New York City Housing and Vacancy

Survey is conducted to understand housing conditions. In this paper, we will focus on the client's income and determine if any factors contribute to a resident's income.

Exploratory Data Analysis

Data

In this NYC housing data, we will analyze a random sample of 299 residents and four variables collected from the New York City Housing and Vacancy Survey. We will examine the relationship between client income and three explanatory variables: age, maintenance deficiencies, NYCMove year.

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 first first few lines of data are:

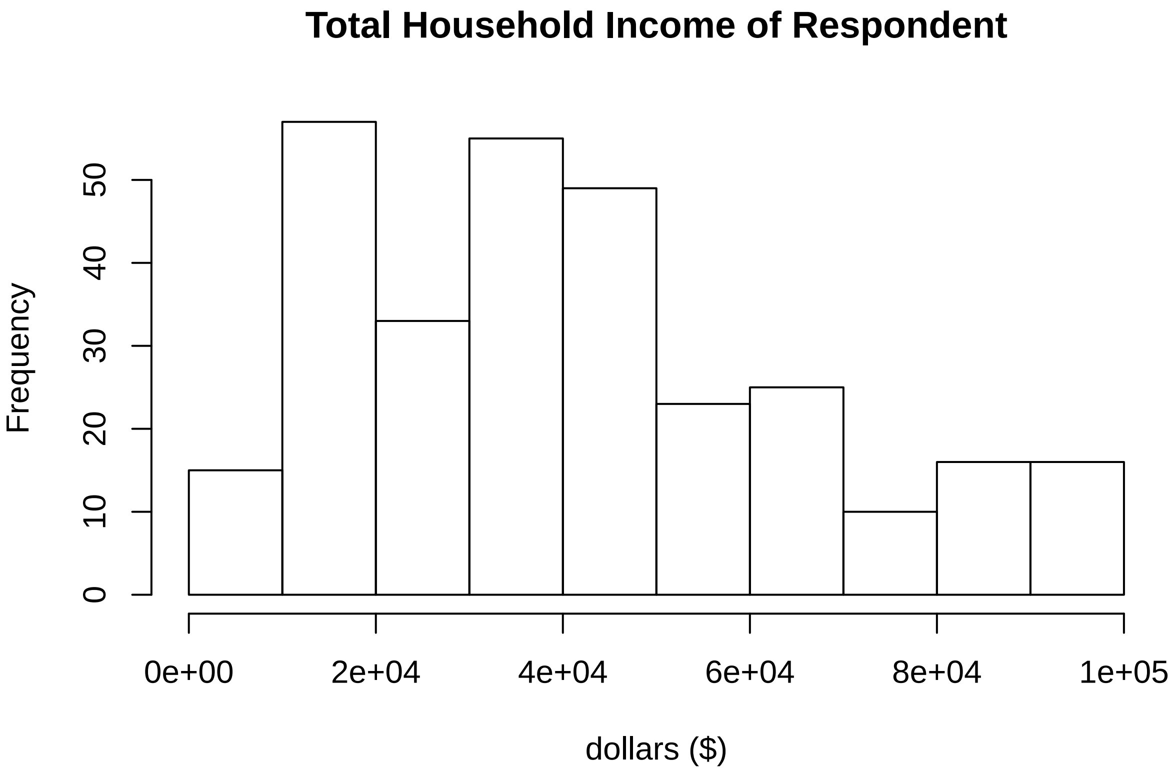
```
head(nyc)
```

```
## # 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
```

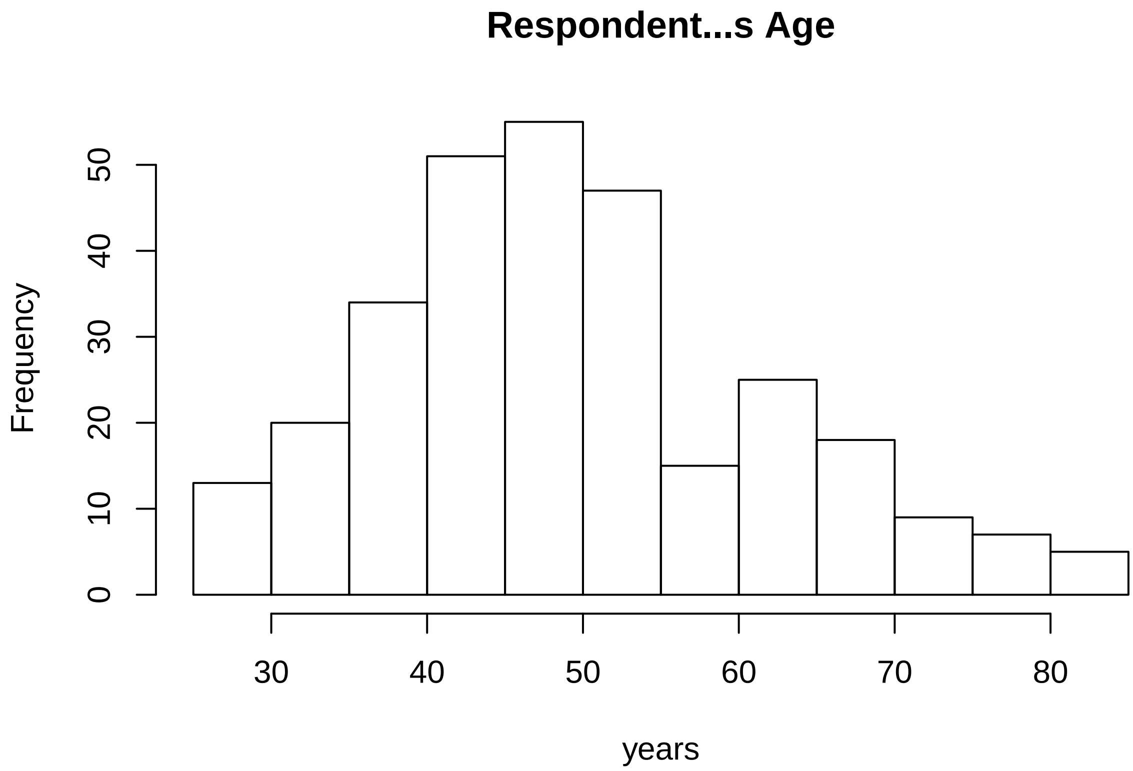
Univariate EDA

As a first step in the analysis, we will explore each variable individually. We will use histograms to look at the distribution of the four variables.

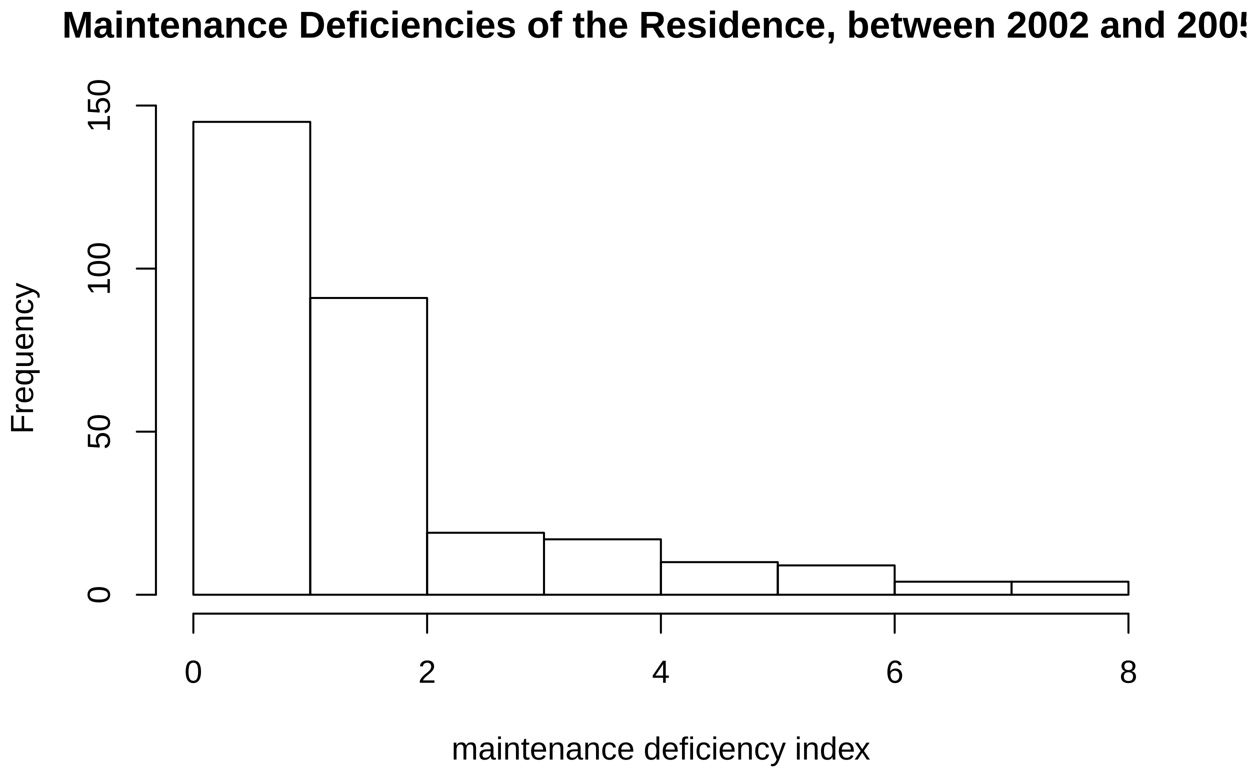
```
hist(nyc$Income,
     main = "Total Household Income of Respondent",
     xlab = "dollars ($)")
```



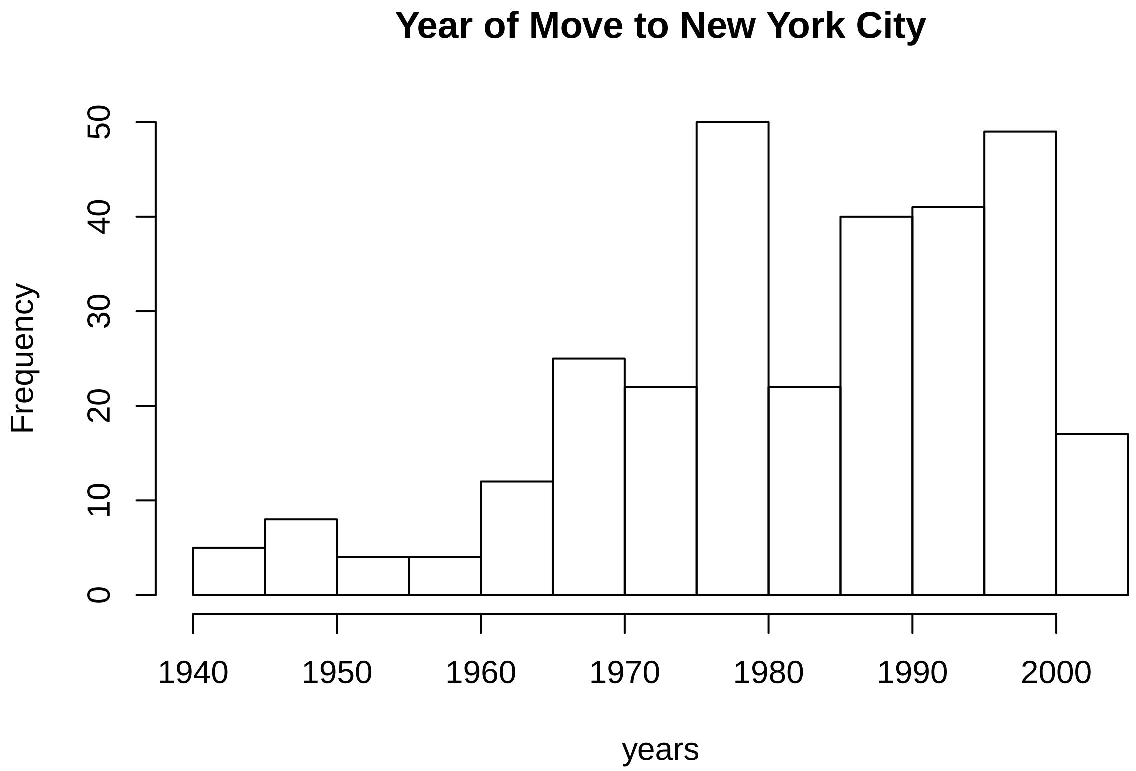
```
hist(nyc$Age,
     main = "Respondent's Age",
     xlab = "years")
```



```
hist(nyc$MaintenanceDef,
     main = "Maintenance Deficiencies of the Residence, between 2002 and 2005",
     xlab = "maintenance deficiency index")
```



```r
hist(nyc$NYCMove,
     main = "Year of Move to New York City",
     xlab = "years")
```



Univariate Graphical Summary

We will find the univariate graphical summary with numerical summaries containing six values.

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 Maintainence 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 NYCMove Year:

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

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

Looking at both the univariate graphs and summary statistics of the variables, the distribution of income could be considered bimodal, trimodal or uniform, but we will need more data to distinguish the distribution. We see spikes in the variable income from 10000-20000 and 30000-50000.

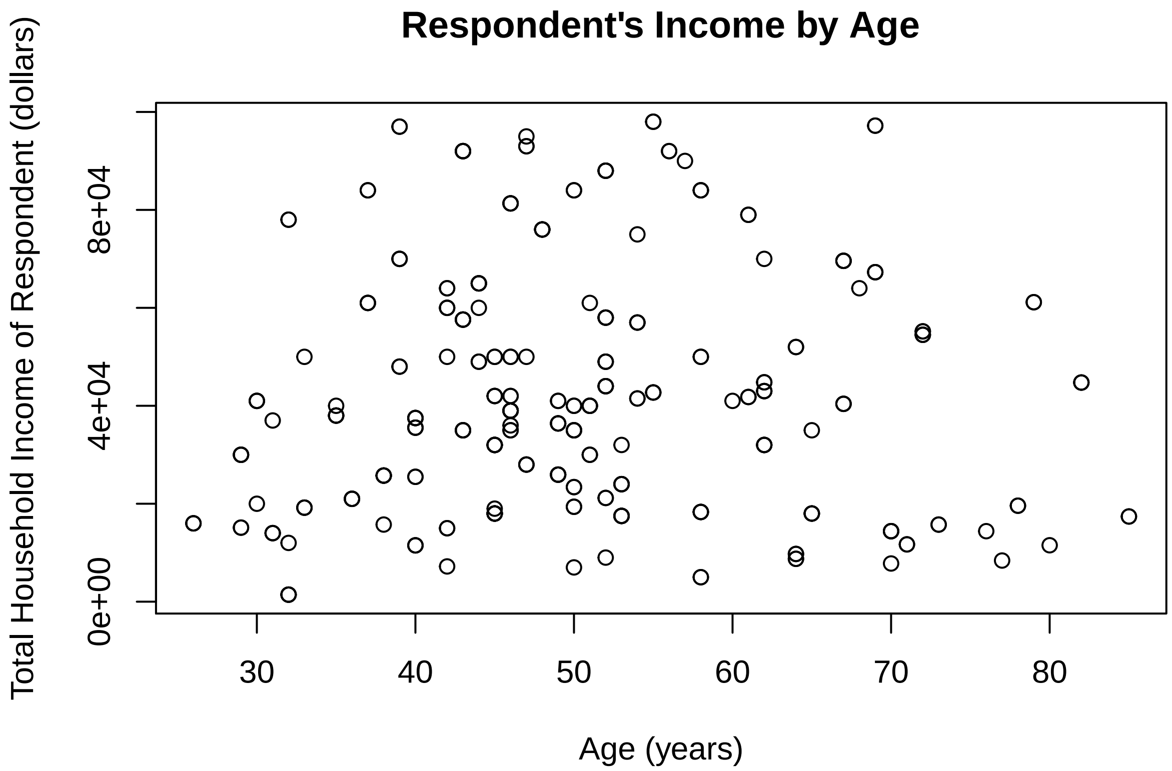
The distribution of age is approximately unimodal with the the mean of 50.03 and median of 49 being very similar. The distribution center is around 50, and the range and IQR are 59 and 16, respectively.

The distribution of maintainence deficiency is skewed right with a range from 0 to 8 with an average of about 1.98. Most of the maintenance deficiencies range from 1 to 2 in the distribution.

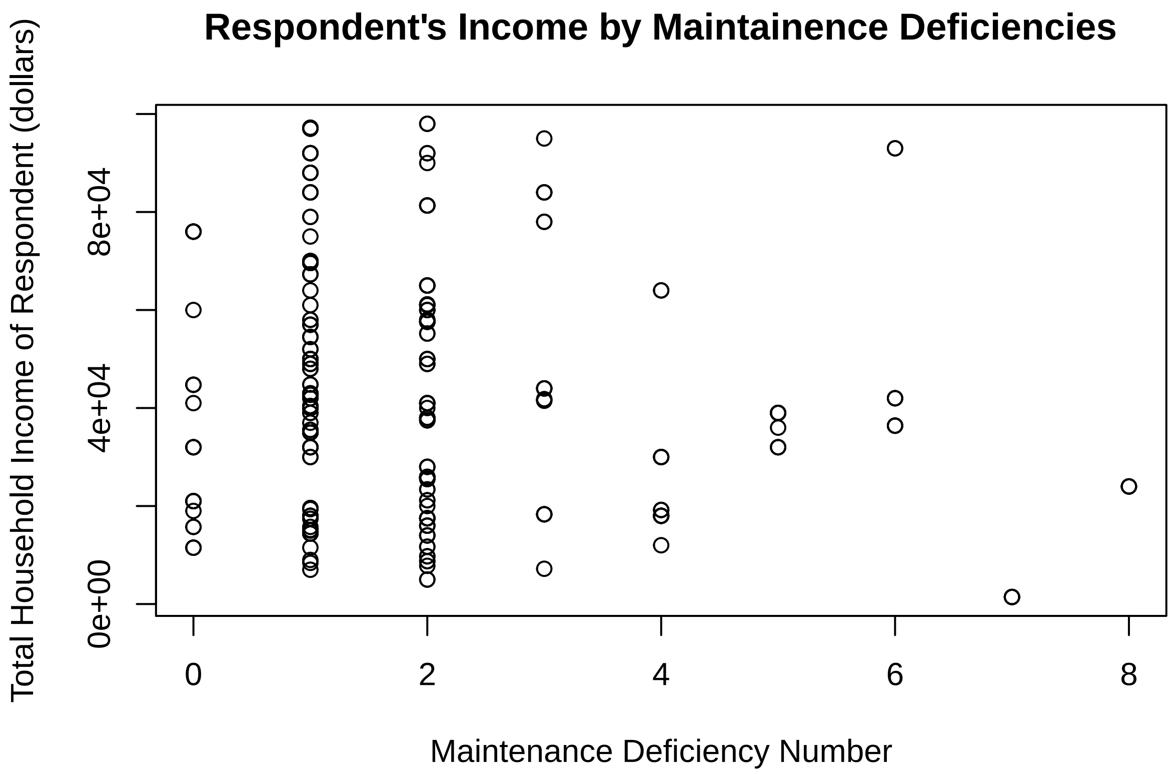
The distribution of the NYC Move year can be considered skewed left and bimodal as well with spikes between 1975-1980 and 1995-2000. The range of the distribution is 1942 to 2004 with the mean and median being very close in values of 2 and 1.98, respectively.

Bivariate EDA

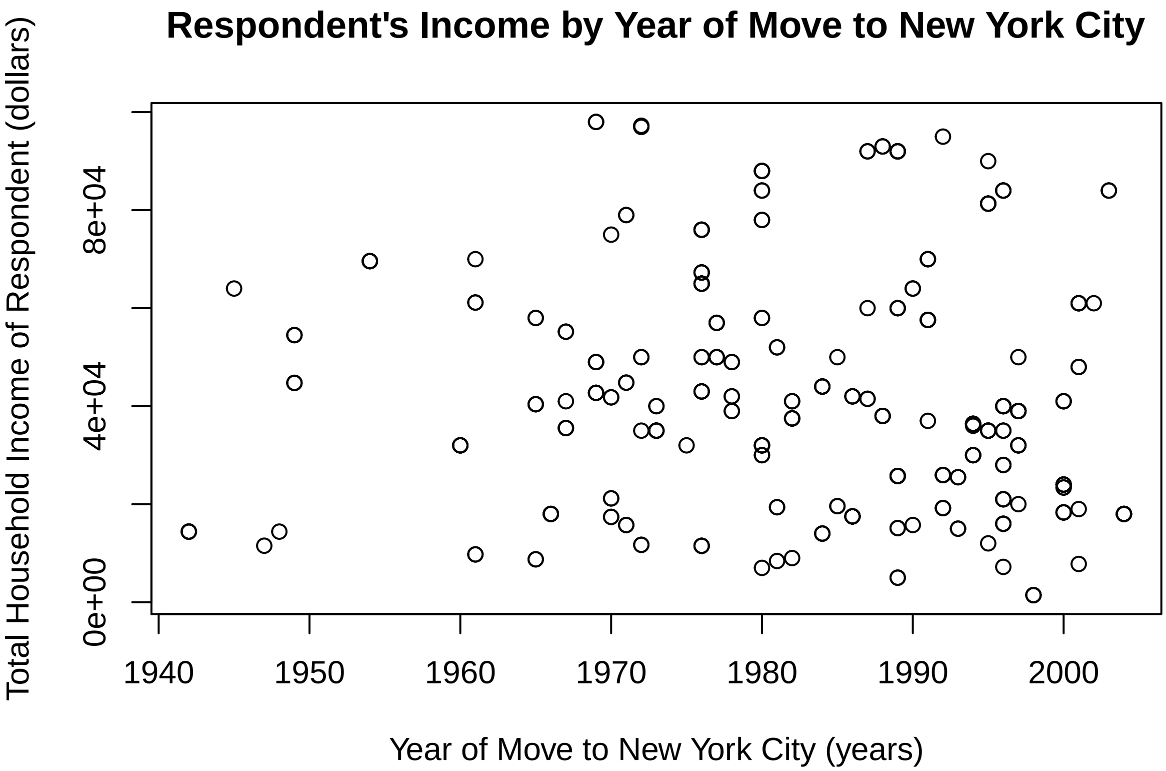
```r
plot(Income ~ Age,
     data = nyc,
     main = "Respondent's Income by Age",
     xlab = "Age (years)",
     ylab = "Total Household Income of Respondent (dollars)")
```



```r
plot(Income ~ MaintenanceDef,
     data = nyc,
     main = "Respondent's Income by Maintainence Deficiencies",
     xlab = "Maintenance Deficiency Number",
     ylab = "Total Household Income of Respondent (dollars)")
```



```
plot(Income ~ NYCMove,
     data = nyc,
     main = "Respondent's Income by Year of Move to New York City",
     xlab = "Year of Move to New York City (years)",
     ylab = "Total Household Income of Respondent (dollars)")
```



From the analysis of our graphs, we find that the total household income of residents is not linearly associated with age, maintenance deficiencies, and year of the move to New York City. There is no clear association between the variables in regards to direction and linearity. Even though the association between income and year of move shows a slight negative association, it is not very strong.

Modeling

Transformations

Learning about the linear relationships among our variables, we start building linear regression models to predict the resident's total income. To start, we look at the histogram of our response variable, which isn't symmetrical, showing that a transformation might be needed. We will look at potential transformations in the response variable by taking the logarithm of the variables.

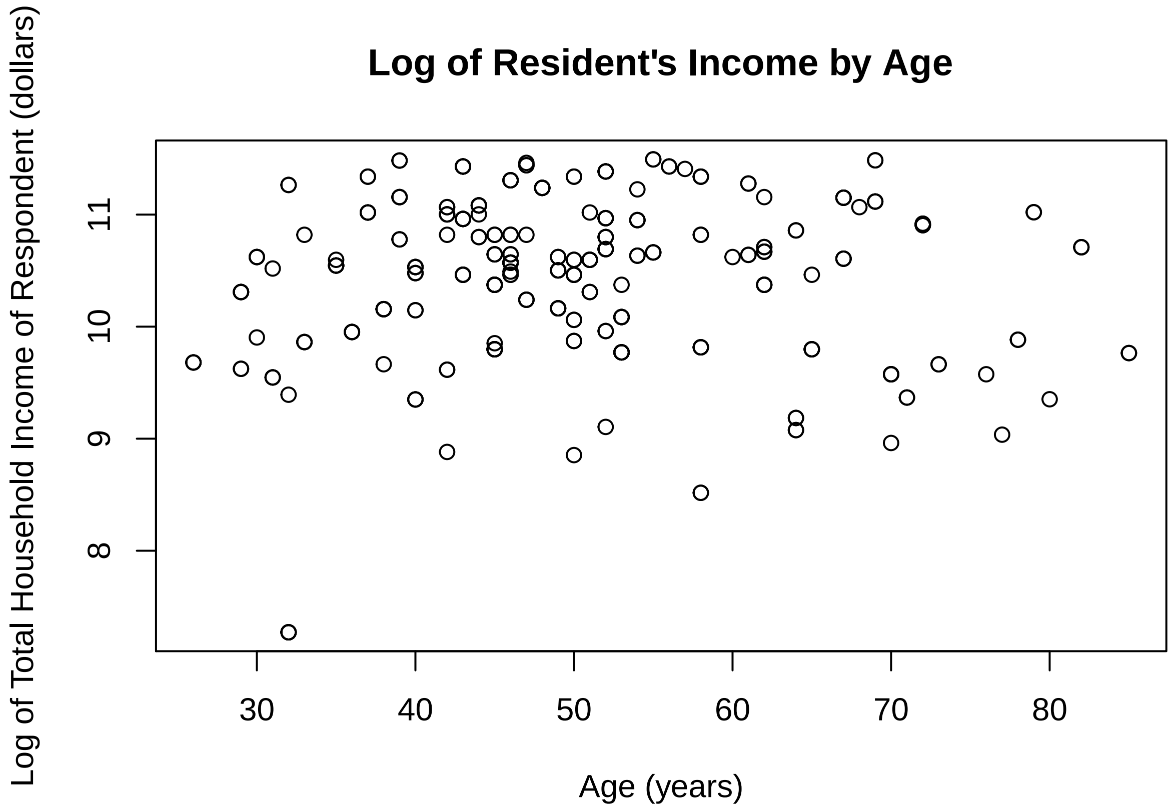
```
min(nyc$Income)
```

```
## [1] 1440
```

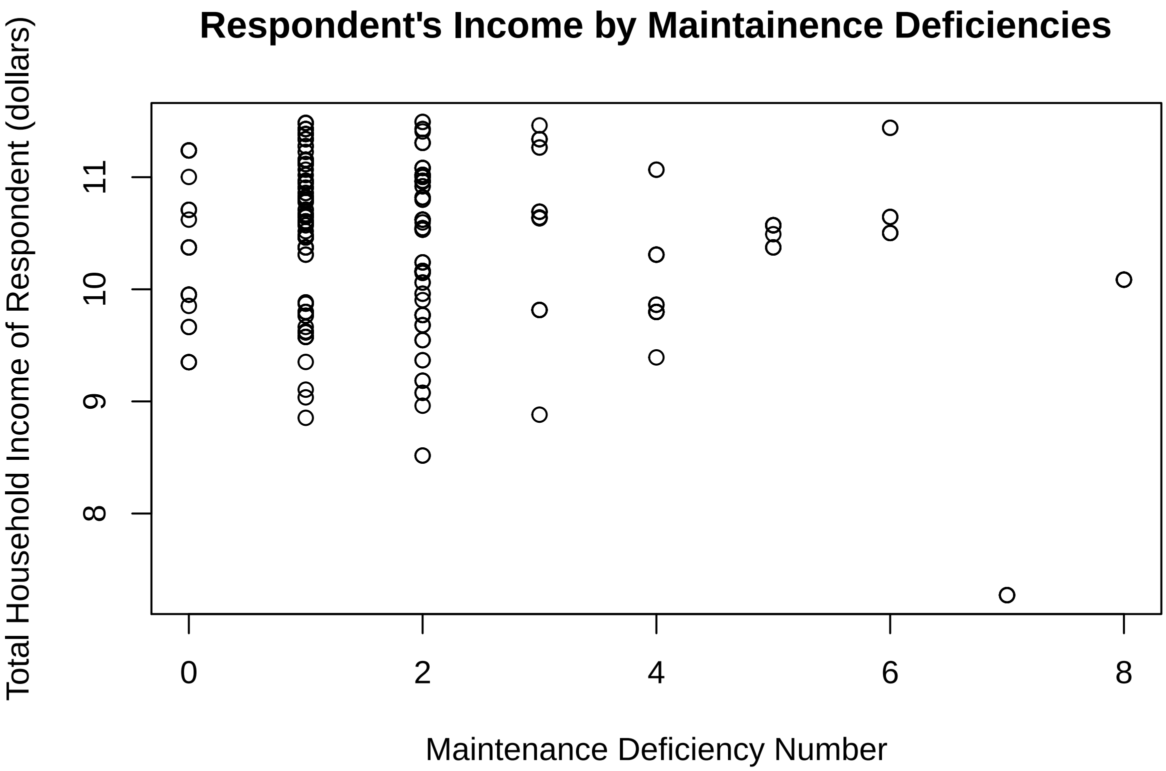
The minimum is greater than 1, so we can take the logarithm of the variable Income.

```
nyc$log.Income <- log(nyc$Income)
nyc$log.Age <- NULL
nyc$log.MaintenanceDef <- NULL
nyc$log.NYCMove <- NULL
```

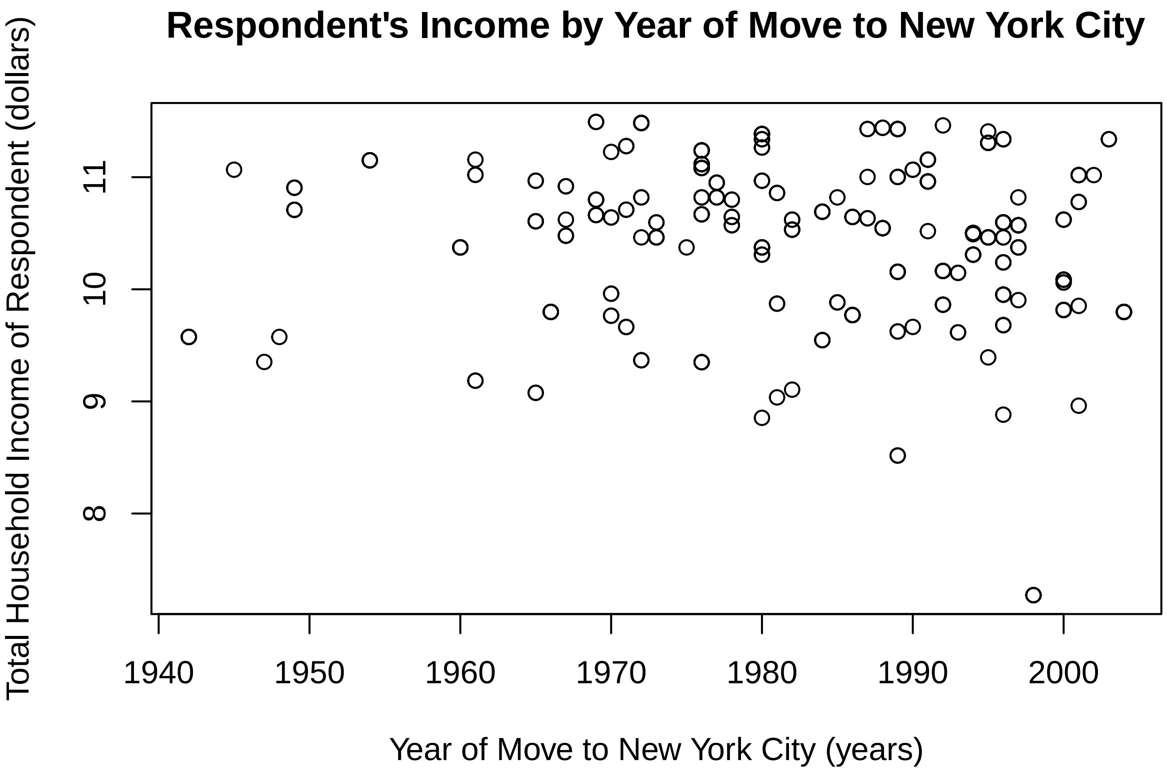
```
plot(log.Income ~ Age,
     data = nyc,
     main = "Log of Resident's Income by Age",
     xlab = "Age (years)",
     ylab = "Log of Total Household Income of Respondent (dollars)")
```



```
plot(log.Income ~ MaintenanceDef,
     data = nyc,
     main = "Respondent's Income by Maintainence Deficiencies",
     xlab = "Maintenance Deficiency Number",
     ylab = "Total Household Income of Respondent (dollars)")
```



```
plot(log.Income ~ NYCMove,
     data = nyc,
     main = "Respondent's Income by Year of Move to New York City",
     xlab = "Year of Move to New York City (years)",
     ylab = "Total Household Income of Respondent (dollars)")
```

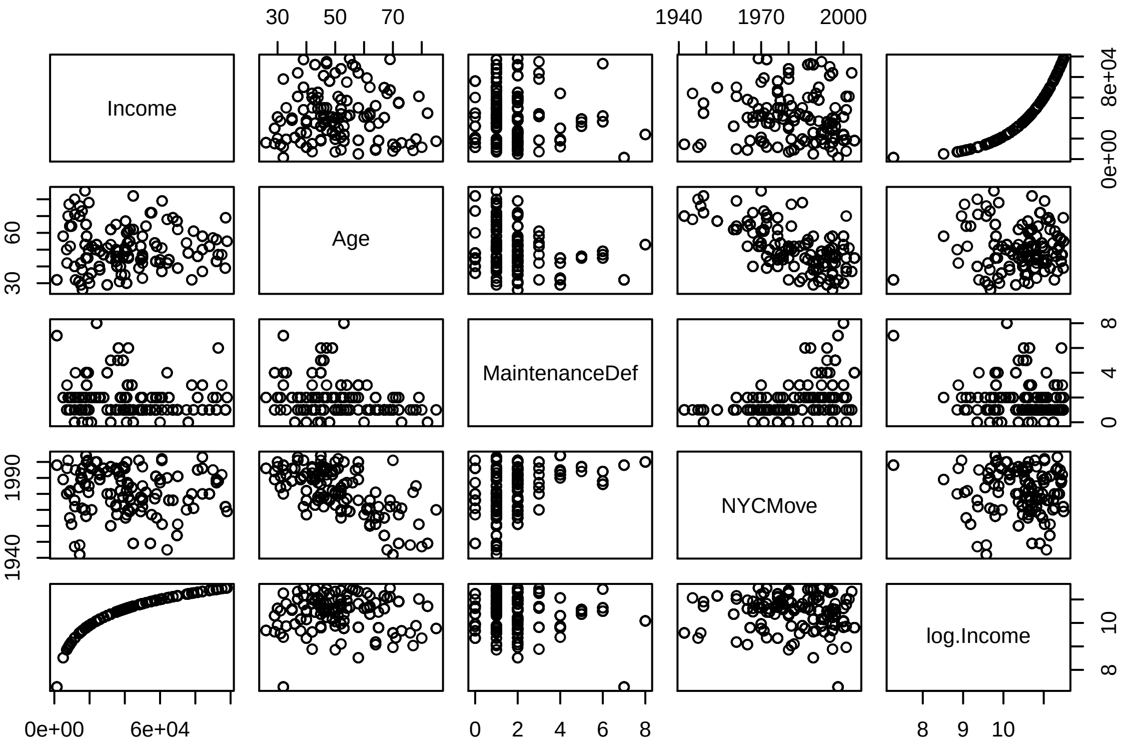


From the analysis of our graphs, we find that the respondent's total household income is slightly linearly associated with age, maintenance deficiencies, and year of the move to New York City, showing a slight negative association between income and the other three variables.

In our transformed bivariate exploratory data analysis, we discovered that the variables have either a weak or strong relationship with resident's income. Because the variables may be useful for this model, we'll need to check for multicollinearity. We check the pairs plot to find any strong correlations between explanatory variables.

Pairs Plot

```
pairs(nyc)
```



There is a slightly strong linear relationship between age and NYCMove, so we should have concerns about multicollinearity and their effect on the model.

Because of this concern, we will check the variation inflation factors (vif) for these explanatory variables to check for multicollinearity.

VIF

```
nyc.full.mod <- lm(Income ~
                     Age + MaintenanceDef + NYCMove,
                   data = nyc)
vif(nyc.full.mod)
```

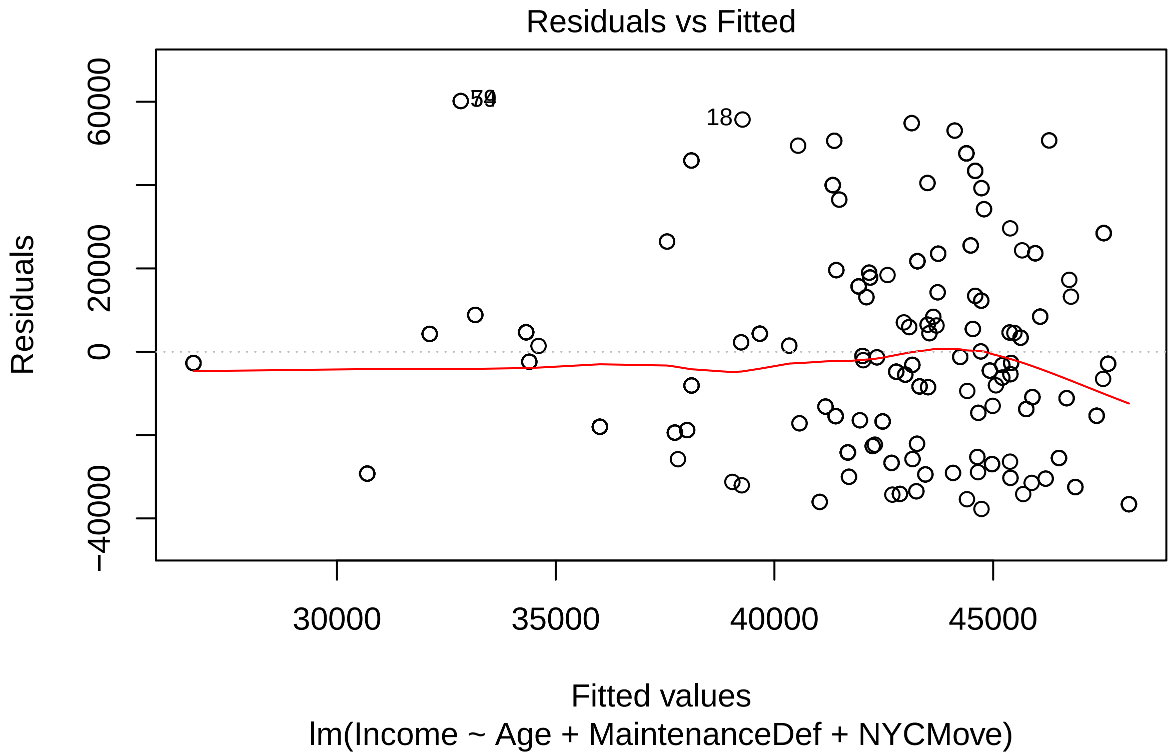
```
##            Age MaintenanceDef        NYCMove
##       1.687649       1.267728       1.999724
```

Because none of the variables have a vif over 2.5, our model does not have a potential multicollinearity issue.

Residual Plot

After confirming that we have a model without a multicollinearity issue, we will create residual diagnostic plots with the explanatory variables.

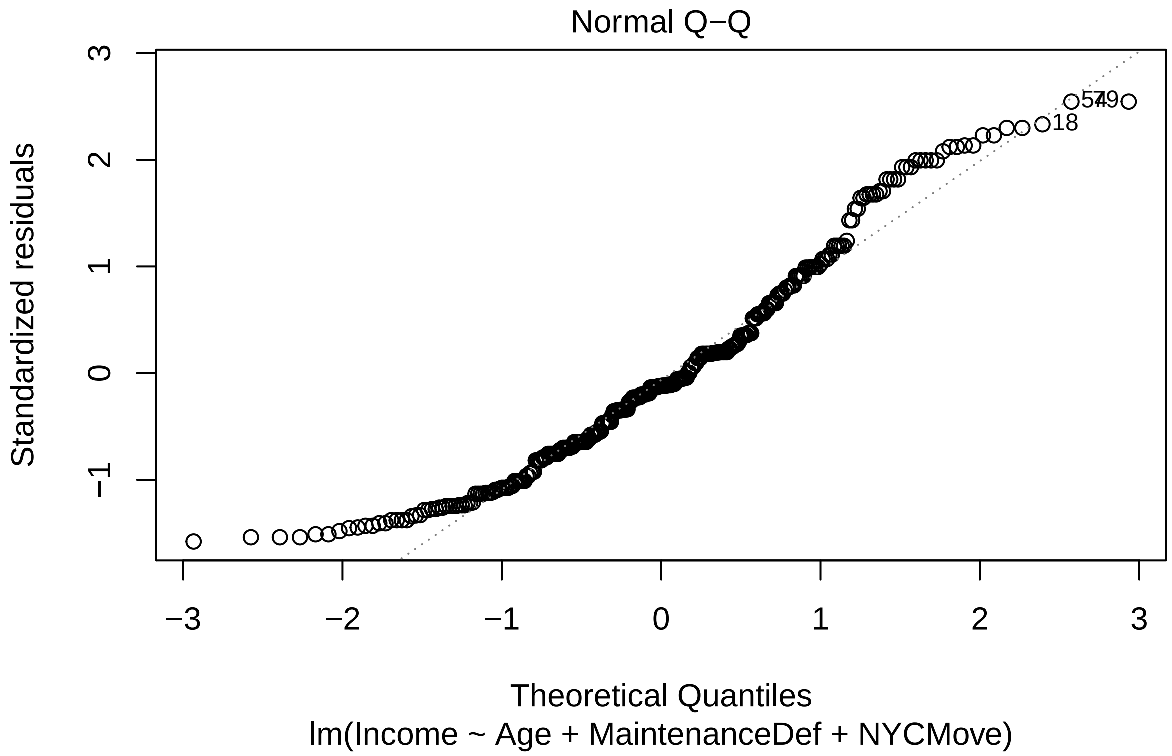
```
plot(nyc.full.mod,
     which = 1)
```



On the residual plot, we discover that aside from a couple of points, like rows 18, 54, and 79, having large residual values, the independence, mean zero, and spread assumptions are reasonably justified from the information we got, which presents no pattern and a roughly equal spread.

Q-Q Plot

```r
plot(nyc.full.mod,
     which = 2)
```



On the Q-Q plot, we can see the values curving away from the line at the lower and curving towards the line in the upper end away from the cluster of values. Unlike the residual plot, rows 18, 54, and 79 are closer to normality. Even though there are many points away from the others, they do not appear severe enough to invalidate the normality condition because all of the other points are close to the line on the plot. This was also around the best normality diagnostic with the models before the transformation.

Regression Analysis Summary

So, we will produce the regression analysis summary for the final model.

```r
summary(nyc.full.mod)
```

```
##
## 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
```

After trying both the original model and the transformed model, the original model can still be considered a reasonable model; even though the R^2 value is larger for the transformed model, the original p-value still shows that the predictors are significant in the data. Also, with no multicollinearity issues due to the low vif scores, the original model is still reasonable to consider for analysis.

We will consider that this a reasonable model to predict resident's Income because the explanatory variables showed a relatively linear relationship with Income, which justifies the linearity of the multiple linear regression model. This model produced an R^2 of 0.02981, which is slightly less than the transformed model but still just as stable.

However, the model is still statistically significant because the regression F-test p-value is 0.03005, which is less than 0.05. Additionally, each coefficient is significant in the model.

We see negative coefficient values associated with Age, MaintenanceDef, and NYCMove, which lines up with EDA results.

Overall, this linear regression model does not have a potential multicollinearity issue, has negative coefficients consistent with the EDA, and has a reasonable R^2 value while balancing the residual diagnostics and relative simplicity.

We are confident that young age, fewer maintenance deficiencies, and residents who moved to New York City a long time ago are associated with high Income.

Prediction

After finding a model that satisfies the assumptions, we are interested in predicting the income for a residence with three maintenance deficiencies with a resident of 53 years of age who moved to New York City in 1987.

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

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

The predicted income of the 53-year-old resident who moved to New York City in 1987 and has a residence with three maintenance deficiencies is $39320.23.

Discussion

From this data analysis, we discover that the respondent's income is indeed related to the resident's age, the number of maintenance issues, and their New York City move-in year with not strong multicollinearity issues.

In the model, all three predictors are significant.

We note an issue with a couple of outliers, especially seen in the residual plots, so further investigation of this issue is advised.

We also note that information about the type of household the respondent lives in could be useful for the current data because we might want to know if the residence type may also help determine the resident's income.

Overall, the data analysis helps us learn that residents have to have high-paying incomes to live in New York City and maintain good housing conditions. The New York City Housing and Vacancy Survey should continue to give the survey out every three years or even more frequently. It helps us understand the type of residents living in these households and how certain factors of their life have affected the way they live in New York City. These surveys should also add new categories, like the household type, to compare more explanatory variables to the response variable. Analyses of surveys are beneficial to both the current residents of New York City and future residents who may want to learn what income, age, and year is best for them to move to the city.