Predicting Income of NYC Residents

Zaina Saif zsaif

Contents

Introduction

New York City is the United States' largest city and a significant center for culture and finance worldwide. Its high cost of living, especially in housing, has been known to pose economic challenges for its residents. This report investigates the impact of age, maintenance deficiencies, and the year of moving to NYC on household income. By analyzing the data we aim to understand the factors influencing income in NYC, which could help support residents in making better housing choices and enhance our understanding of the city's economic conditions.

Exploratory Data Analysis

Data

The dataset is derived from the New York City Housing and Vacancy Survey, conducted every three years to capture a comprehensive snapshot of the city's housing conditions. In this data we observe a sample of 299 NYC residents grouped by 4 variables. This report takes particular interest in the response variable, Income, in comparison to the three exploratory variables: Age, MaintencanceDef, and NYCMove. The four key variables are described as follows:

Income: the total household income in dollars

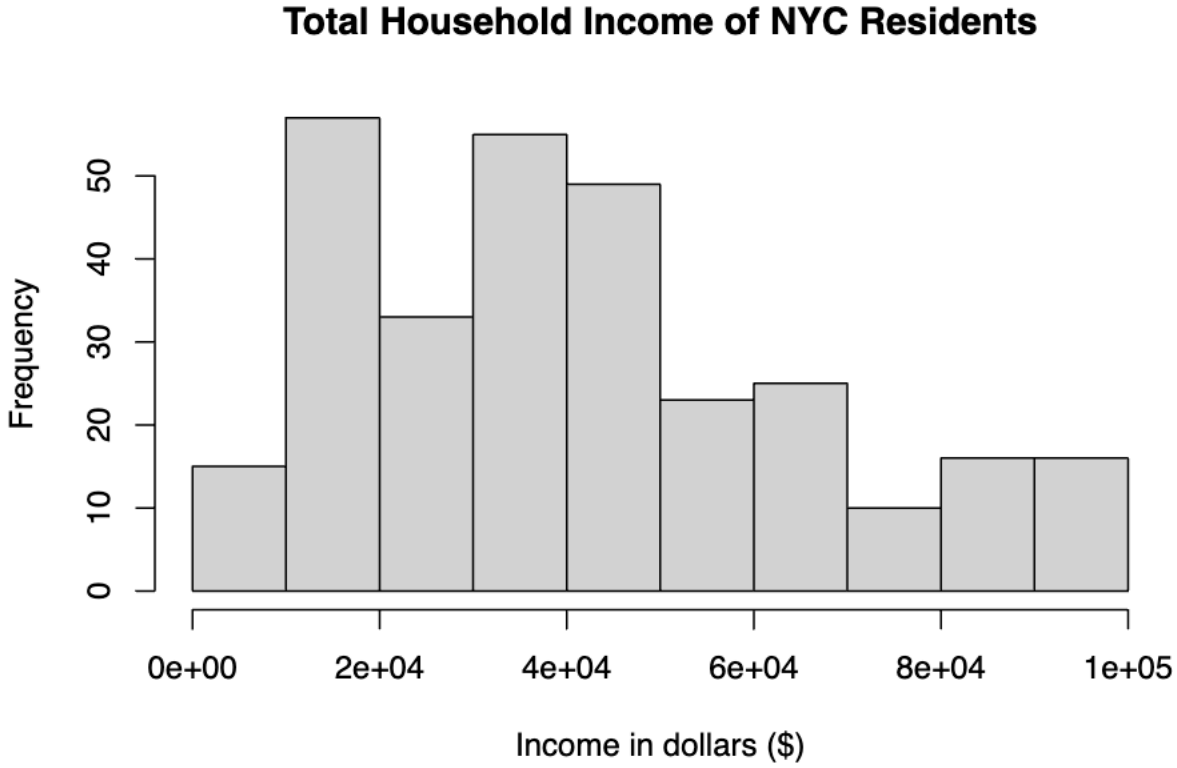
Age: the respondent's age in years

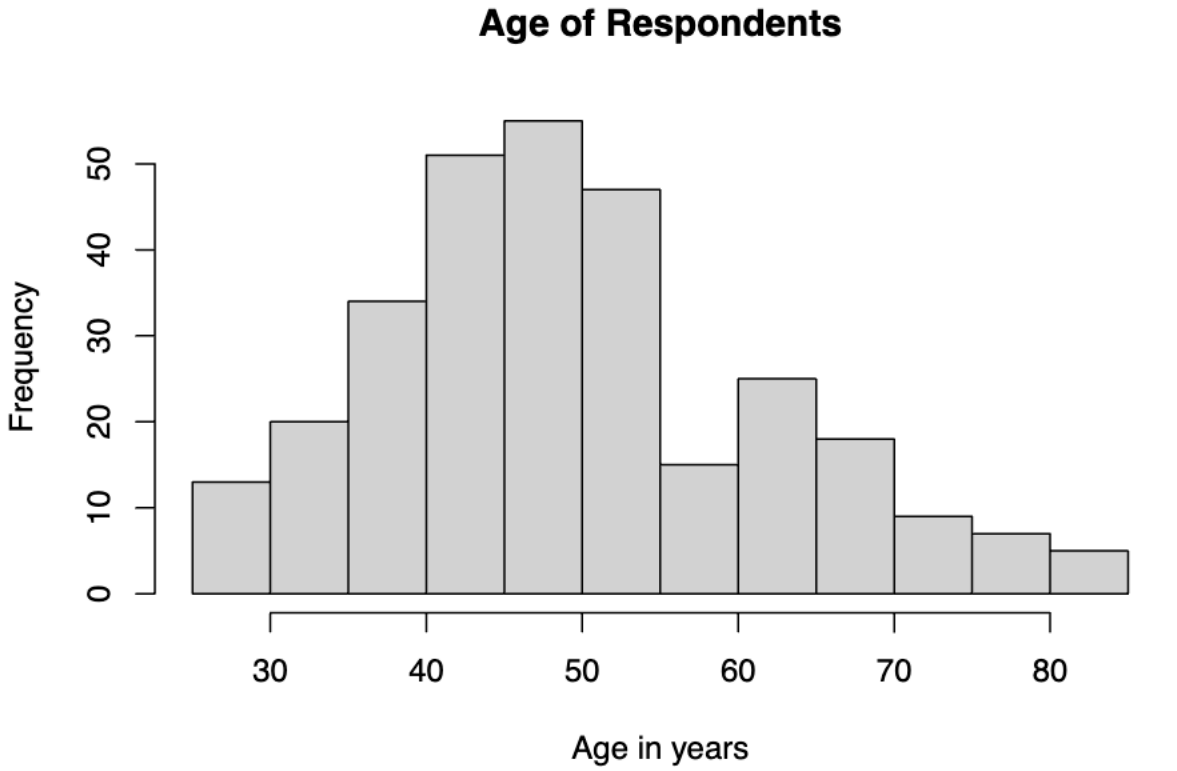
MaintenanceDef: the number of maintenance deficiencies of the residence, recorded between 2002 and 2005

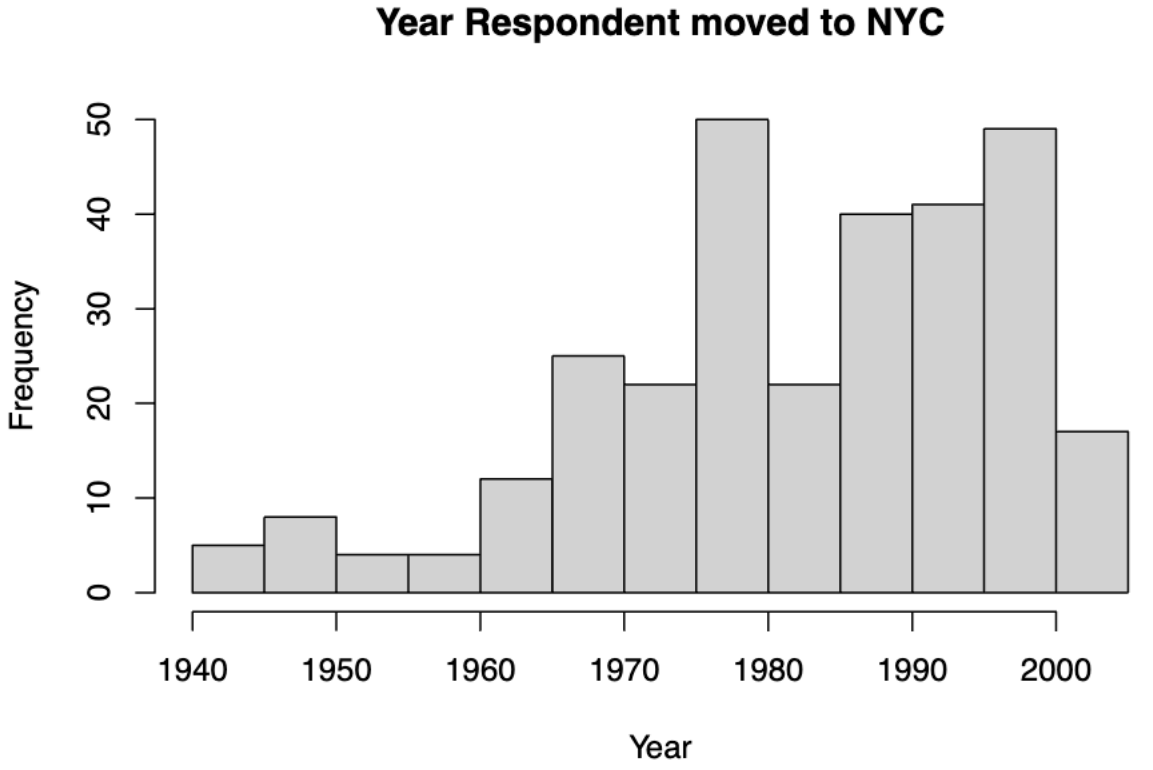
NYCMove: the year the respondent moved to New York City

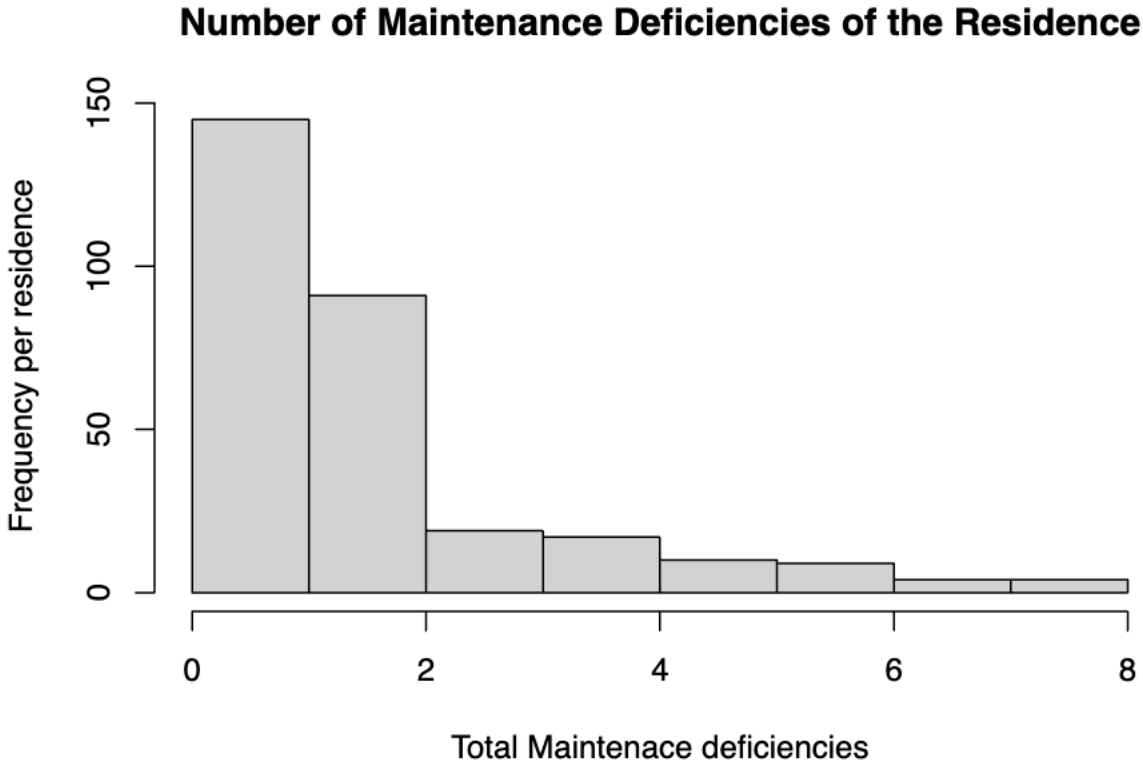
Univariate EDA

By using histograms we can examine each variable individually, examining the distribution of each variable in comparison to the frequency of that response.









The numerical summaries for the univariate data analysis are as follows:

Income

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

```
## [1] 24201.04
```

Age

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

```
## [1] 12.43678
```

MaintenanceDef

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

```
## [1] 1.619802
```

NYCMove

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

```
## [1] 14.13746
```

Analyzing the data through graphs and numerical summaries, we can make several observations. To begin, the distribution of respondent income appears to be bimodal and slightly right-skewed, with peaks around 1x10^4 ($10,000) and 3x10^4 ($30,000). The mean and median values are similar and the standard deviation is approximately 2.4x10^4 ($24,000). Additionally, the age of respondents tends to show a unimodal and symmetric distribution, with the mean and median nearly identical (a difference of 1.003), and a standard deviation of 12.4 years. Further, the histogram for the number of maintenance deficiencies per residence indicates a unimodal distribution that is skewed right, showing most residences have few deficiencies. The median and mean are very close, at 2.00 and 1.98, and the standard deviation is 1.62. Finally, the distribution for the year the residents moved to NYC is skewed left and bimodial. The mean and median are fairly close with a two year difference and there is a standard deviation of 14.14.

Bivariate EDA

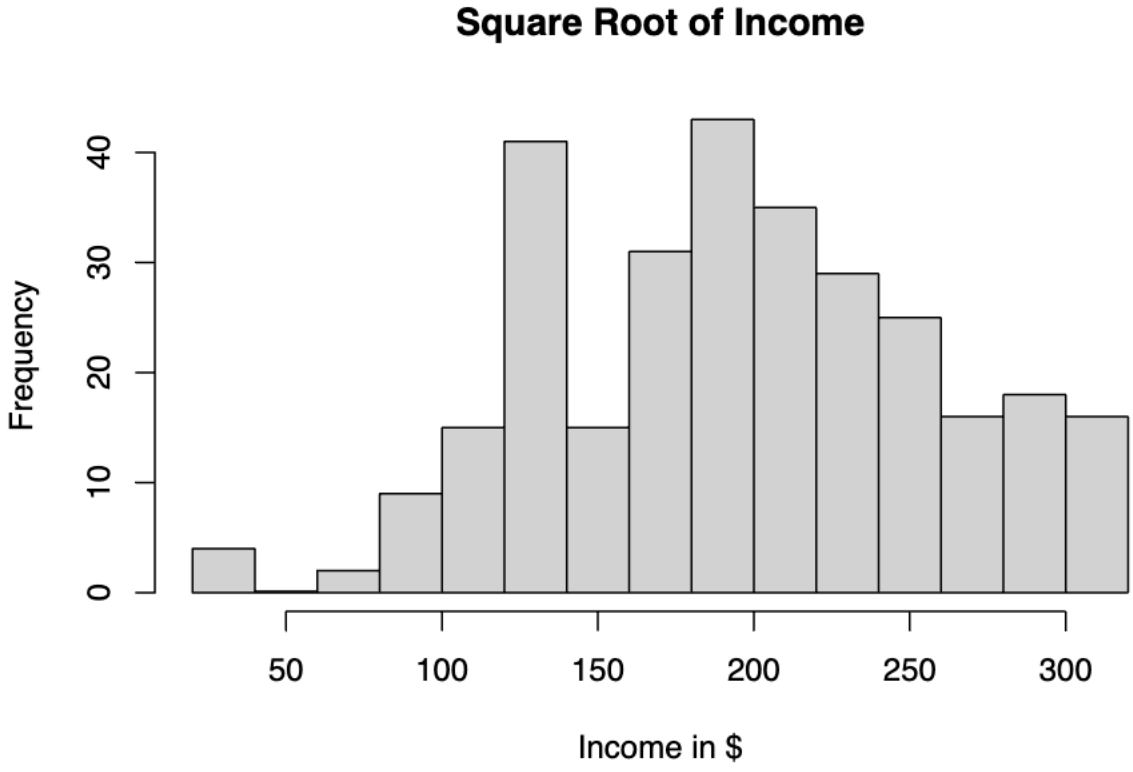
Now we will graphically observe how each predictor relates to the response variable, Household Income. Since income is skewed, we must perform a transformation on the data.

```
nyc$incomeLog <- log(nyc$Income)
hist(nyc$incomeLog)
```

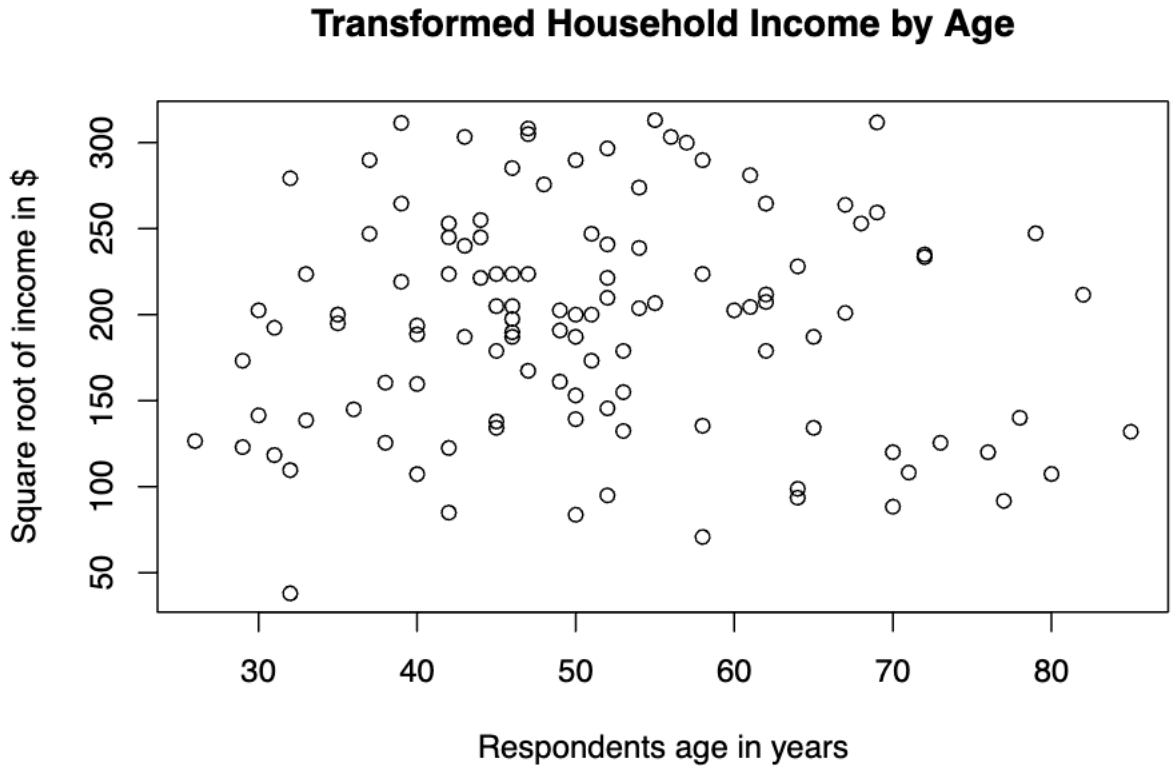
Attempting a log transformation, we observed that the data became more left skewed instead of symmetric. We will use the square root of income as the transformed income variable to make the distribution more symmetric and allowing us to assume normality.

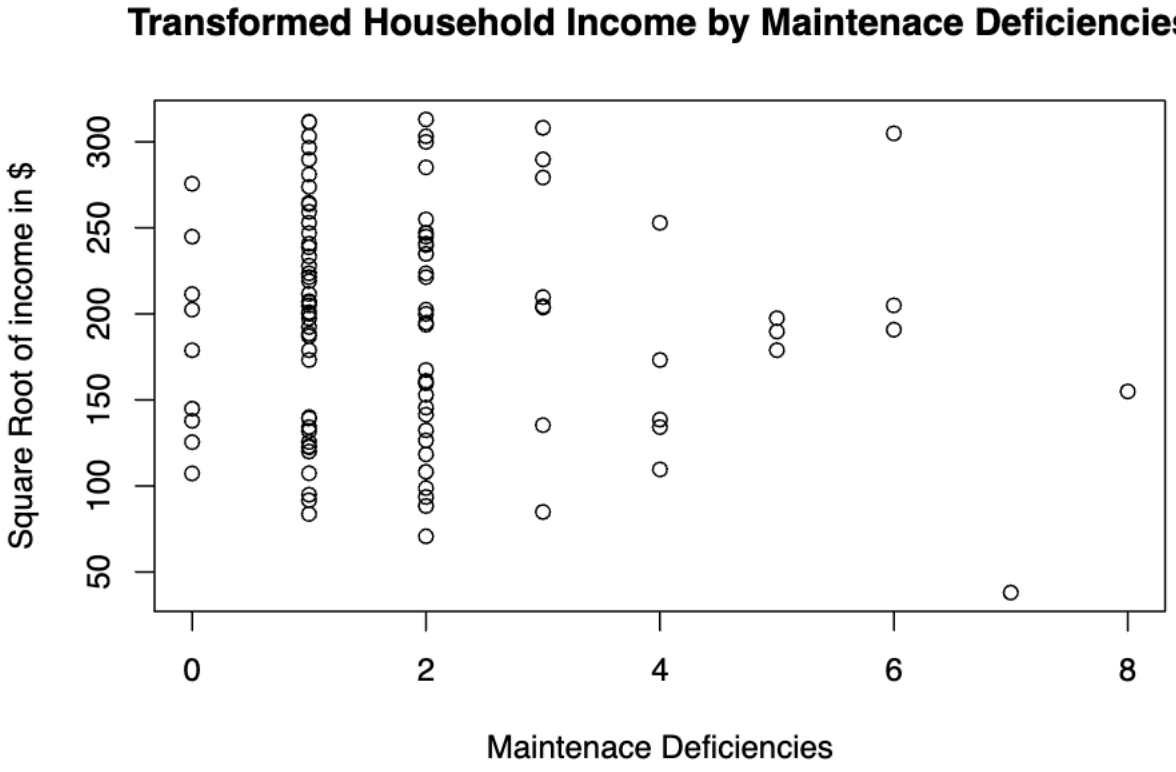
The computed log and histogram of the Income variable, incomeSqd, is below.

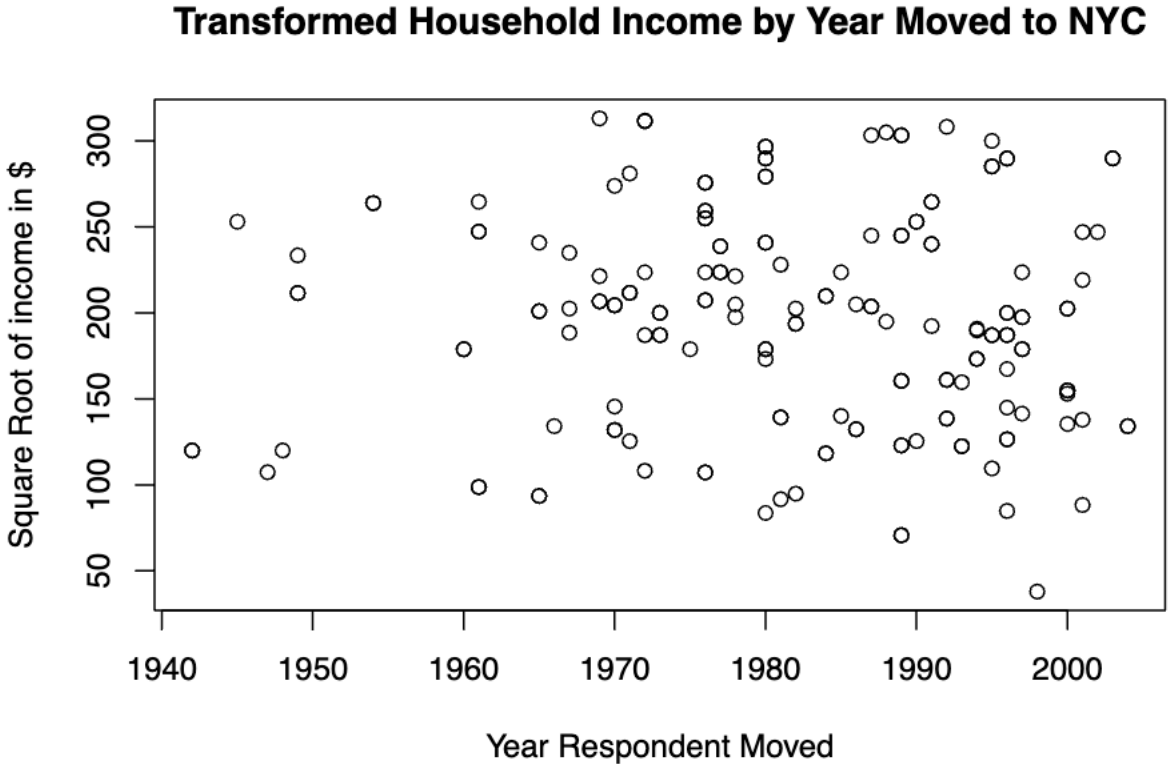
```
nyc$incomeSqd <- (nyc$Income)^(0.5)
hist(nyc$incomeSqd, xlab = "Income in $", main = "Square Root of Income")
```



Now we can compare this transformed income variable with the response variables age, maintenanceDef, and NYCMove.







From observing the scatterplots of the transformed Household Income versus the three response variables, we can make observations about the trend and association.

Income vs Year Moved to NYC

The first plot shows a dispersed set of points without a clear linear trend. Data points appear to be distributed

rather evenly across the years. There is no discernible positive or negative trend. Incomes seem neither to increase nor decrease consistently with the year that respondents moved to NYC. The lack of a clear pattern suggests a weak or non-existent linear association between the year of moving to NYC and log-transformed household income.

Income vs Number of Maintenance Deficiencies

On this plot as well, the data points do not follow a straight line. There's a hint that higher maintenance deficiencies might correspond to lower income, as seen by a slight clustering of lower income values at higher deficiency levels, but this is not a strong or clear trend. There appears to be a weak negative association, given the possible decrease in income with an increase in deficiencies. However, the spread of the data points indicates that the strength of this relationship is not strong.

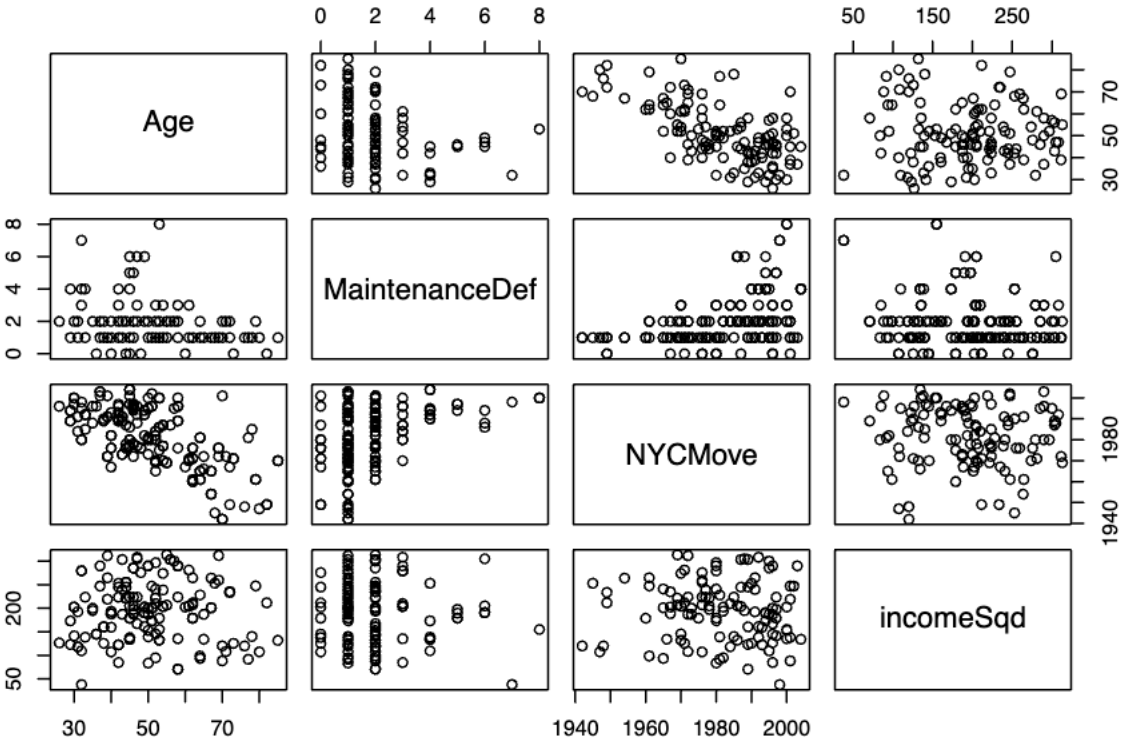
Income by Age

This scatter plot suggests a weakly positive relationship; as age increases, log income appears to increase slightly as well. There is a weak positive trend observable, where log income tends to be higher for middle-aged respondents than for younger or older ones.

Modeling

Now that we have performed data analysis on the variables and their relationships, we can begin to build a linear regression model to predict household income. Looking back at the histogram of the transformed response, incomeSqd, we saw that it is symmetrical. We also saw in the bivariate analysis all variables have some sort of association/relationship with incomeSqd. Although not strong associations, we can continue with these variables in our model before checks for multicollinearity and further motivations to drop any variables.

We produce a pairs plot to check for possible multicollinearity by looking for strong correlation between explanatory variables.



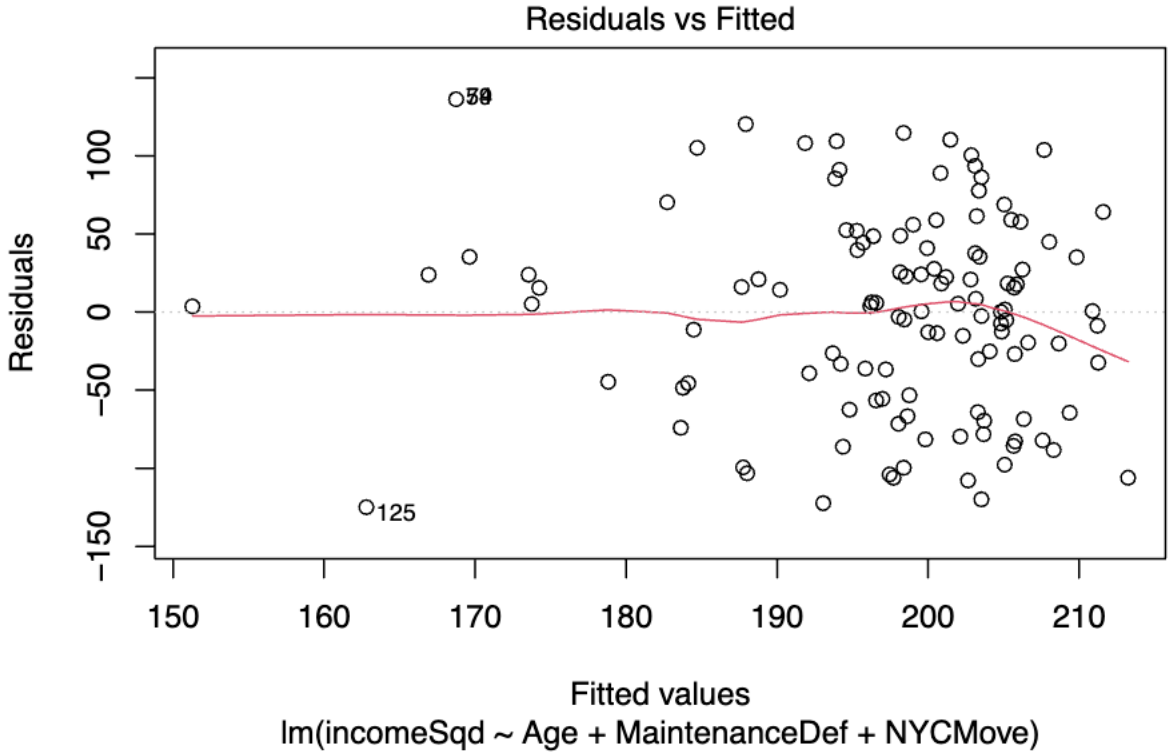
Given the plots, there does not seem to be an immediate concern for multicollinearity based on visual inspection, but we will produce a quantitative check with VIF for definitive assessment.

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

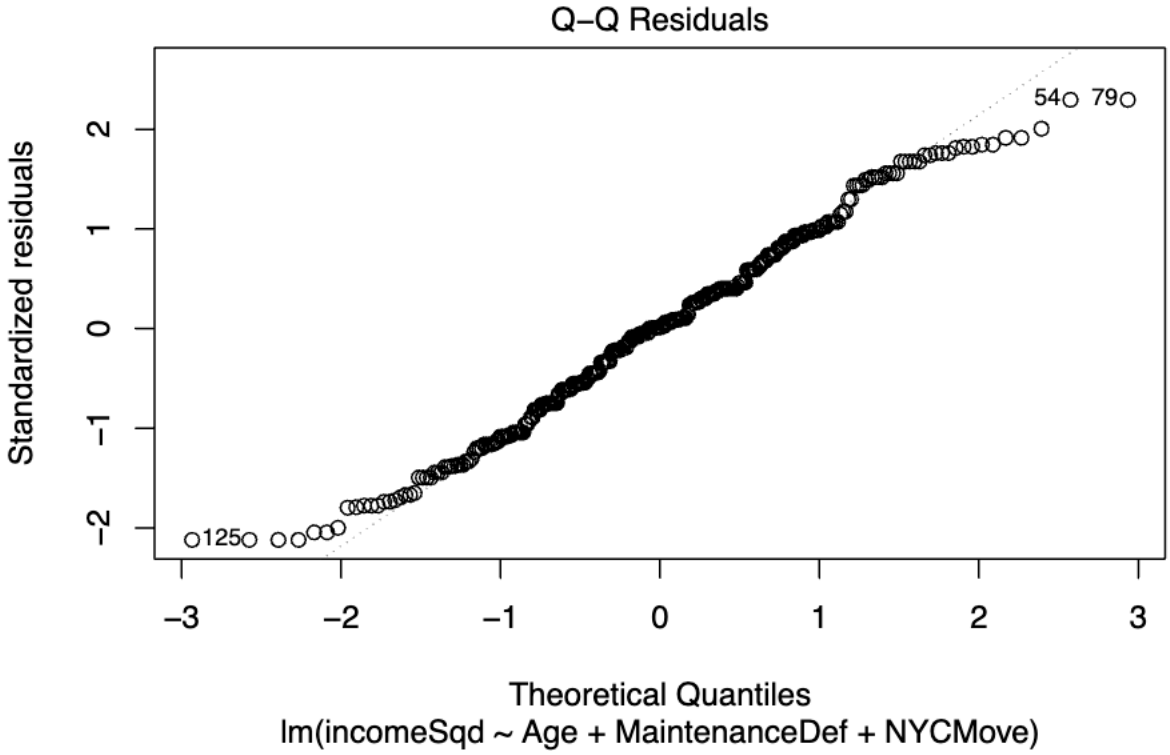
Generally, a VIF value above 2.5 indicates a level of multicollinearity that could be problematic in the regression analysis. Since all the VIF values in the model are below the threshold, we can conclude that multicollinearity is not a concern for the dataset with these three predictors.

Now we compute the residual and QQ plots to observe error assumptions.

The residual plot for the regression model shows a relatively random dispersion of residuals, suggesting that the assumptions of independence and mean of zero are reasonably met. A few potential outliers are present, particularly one with a significant positive residual, which may warrant further investigation. The constant spread assumption holds for the most part, although there's a hint of increasing spread in the residuals as the fitted values get larger. Overall, the residuals are roughly symmetrically distributed above and below the zero line, and present no severe patterns.



The Q-Q plot of the residuals suggests that the normality assumption is generally met, as the majority of points closely follow the line, especially around the center of the distribution. While there are deviations in the tails and a few outliers, these do not appear severe enough to invalidate the normality condition. Considering the various models attempted, this Q-Q plot represents the best normality diagnostic achieved.



We can now produce a regression analysis summary from fitting our model.

```
##
## Call:
## lm(formula = incomeSqd ~ Age + MaintenanceDef + NYCMove, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.852  -44.651    0.651   42.603  136.212
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    685.5447   699.8086   0.980  0.32808
## Age             -0.2070     0.3637  -0.569  0.56966
## MaintenanceDef  -6.7052     2.4203  -2.770  0.00595 **
## NYCMove         -0.2348     0.3483  -0.674  0.50069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.11 on 295 degrees of freedom
## Multiple R-squared:  0.03864,    Adjusted R-squared:  0.02886
## F-statistic: 3.952 on 3 and 295 DF,  p-value: 0.008712
```

The regression summary shows that the chosen model, with square root of income as the response variable and Age, MaintenanceDef, and NYCMove as predictors, has found a statistically significant relationship between the predictors and the response, as evidenced by the F-test with a p-value of 0.008712. This is less than the conventional alpha level of 0.05, indicating that we can reject the null hypothesis that none of the predictors are significantly related to the square root of income.

Despite this overall significance, not all individual predictors display a strong association with the square root of income at the conventional levels of significance. Age and NYCMove specifically have p-values well

above the 10% threshold, suggesting they do not contribute significantly to the model on their own.

We will attempt to fit the model without these variables before dropping them completely and see how the r^2 value is affected.

```
##
## Call:
## lm(formula = incomeSqd ~ MaintenanceDef, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -125.506  -45.404    1.477   43.783  137.723
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     210.707      5.481  38.441  < 2e-16 ***
## MaintenanceDef   -7.245      2.144  -3.379 0.000824 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.96 on 297 degrees of freedom
## Multiple R-squared:  0.03702,    Adjusted R-squared:  0.03378
## F-statistic: 11.42 on 1 and 297 DF,  p-value: 0.0008245
```

The r^2 value with one variable, maintenanceDef, is slightly lower than with the other variables fit in the model. However this change is insignificant for an already small r^2 value. Given these findings, Age and NYCMove may be justifiably dropped from subsequent models to focus on the more impactful predictor.

Prediction

Based on the model, we can estimate the log of the income for a hypothetical resident with specific characteristics. Let's say we want to predict for a resident who has one maintenance deficiency, is 40 years old, and moved to NYC in 2000. Using the model, the prediction would only take into account maintenanceDef and be calculated as follows:

Income = 210.707 + (-7.245 * MaintenanceDef)

Using R as a calculator we get 203.462 as the square rooted income. To get the actual income we square that result. The predicted income for a resident with one maintenance deficiency is approximately $41,396.79. We remark that this is an average resident income given that it is just 2,000 dollars above the mean household income.

Discussion

Throughout this project, we've investigated the determinants of income in New York City, focusing on factors such as maintenance deficiencies (MaintenanceDef), age, and the year of relocation to the city (NYCMove). Our findings have consistently pointed to the value of transforming the response variable, "Income," to the square root of income. This transformation not only led to an improved R-squared value, indicating a better fit for the model, but also reduced skewness and enhanced the normality of the data, fulfilling key assumptions for the application of linear regression.

Notwithstanding, the study's limitations must be acknowledged. The apparent skewness in the variables MaintenanceDef and NYCMove contributed to discrepancies in the residual plots and linearity, which could suggest underlying complexities not fully captured by the model. Future research could benefit from incorporating additional variables, such as household size or occupation, to paint a more comprehensive

picture of the factors affecting income. Such information could help in refining the predictive power of our model.

In conclusion, the analysis presented provides a valuable examination of income-related factors for residents of New York City, contributing insights that align with our initial aim to reinforce the broader understanding of New York City's economic environment.