Predicting Household Income in NYC

Muhammad Hamzah mhamzah

Due Wed, October 25, at 11:59PM

Contents

```r
library("knitr")
library("pander")
library("readr")
library("magrittr")
library("car")
library("interactions")
library("leaps")
```

Introduction

In the context of the U.S. housing market, New York City stands out as a particularly competitive landscape. As per data from the New York City government, a staggering 1/3rd of renters in the city devote more than 50% of their income to pay rent(https://www.nyc.gov/content/tenantprotection/pages/fast-facts-about-housing-in-nyc). Such figures underscore the escalating importance of data analytics in deciphering the reasons behind the exorbitant housing costs. Within the purview of this statistical project, we will juxtapose household income against three key exploratory variables: age, MaintenanceDef, and NYCmove.

Exploratory Data Analysis

Data

In this data set, we analyze a set of 299 responses from the **"New York City Housing and Vacancy Survey (NYCHVS)"**, which is organized every three years in an attempt to understand the housing situation in NYC, and the interplay/effect of the four variables involved.

Due to our interest in understanding what factors affect the household income, we examine the relationship between Household Income and 3 explanatory variables: Age, MaintenanceDef, NYCMove

We summarize the variables as follows:

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.

Income: Total household income (in $) [the response variable].

We will have a look at the first first few lines of data appear in order to have a better idea of the structure of the data we are dealing with

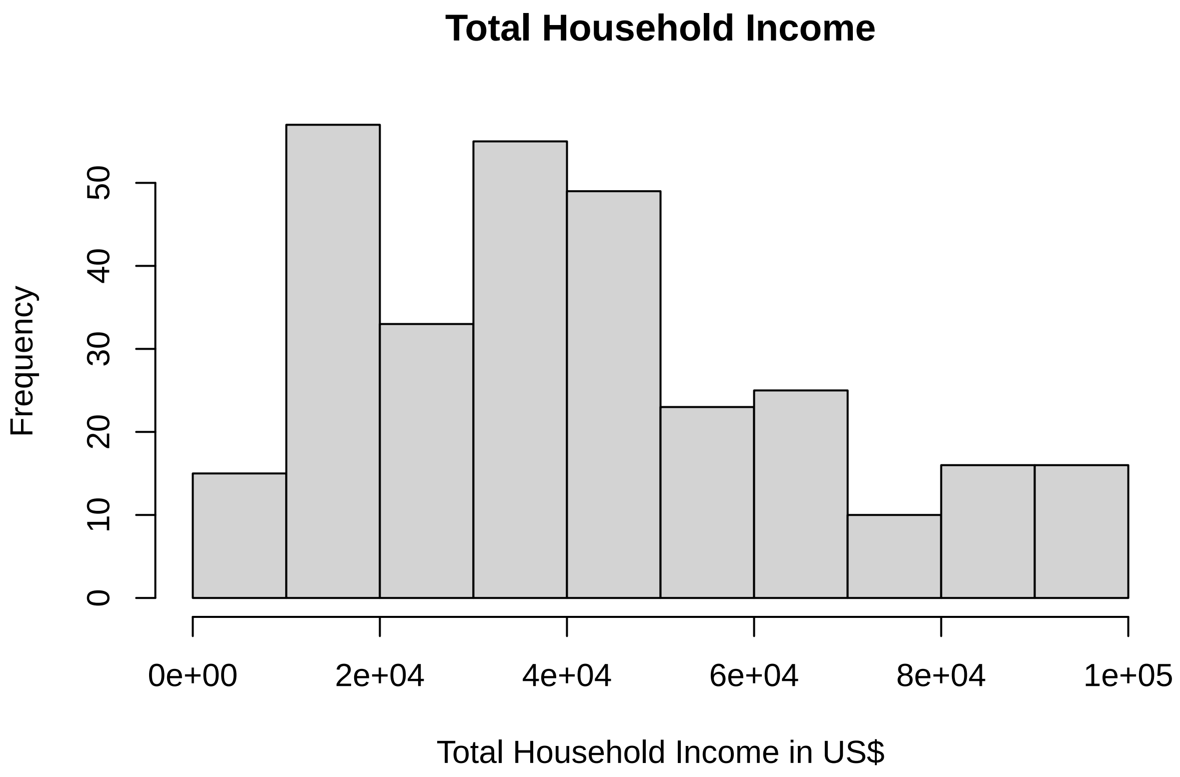
```
head(nyc)
```

```
## # 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
```

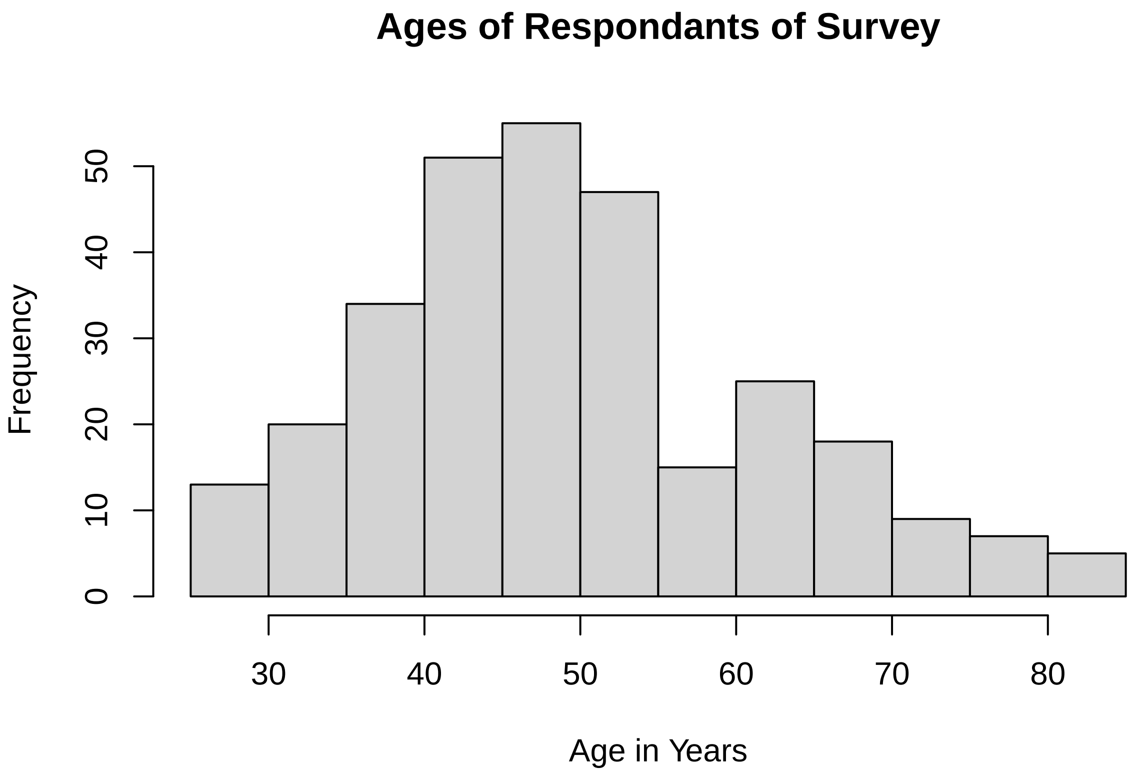
Univariate Exploration

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

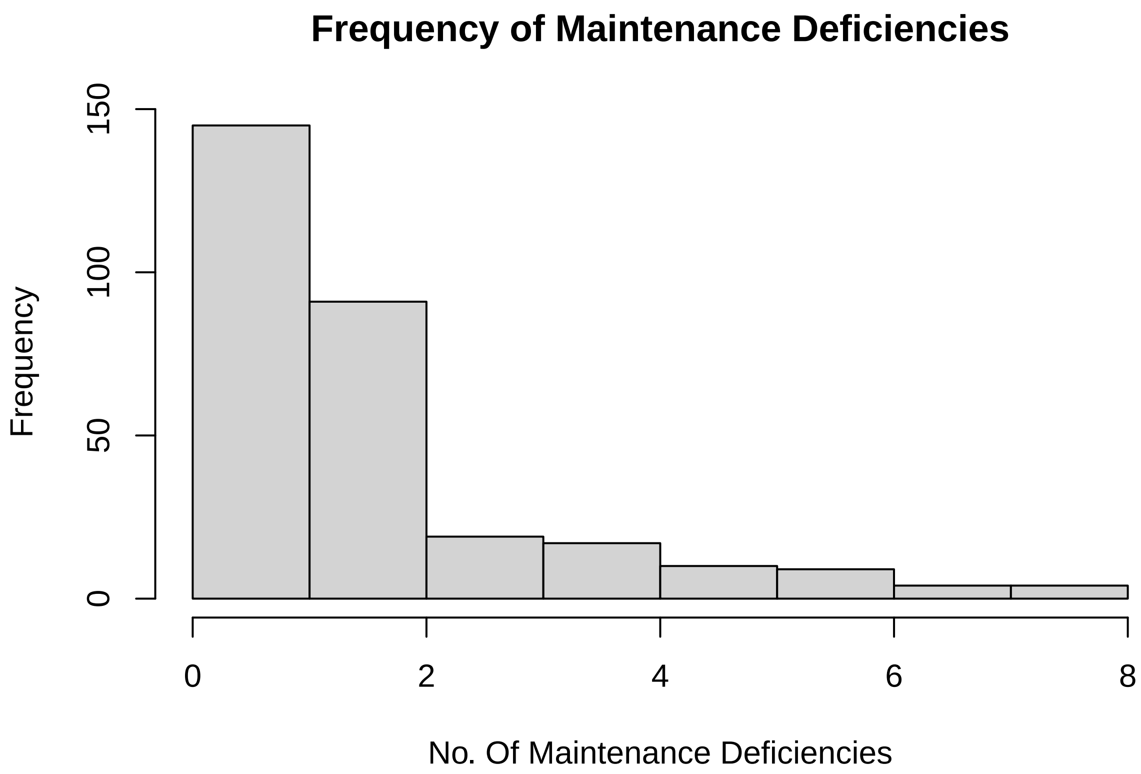
```
hist(nyc$Income,
     main= 'Total Household Income',
     xlab='Total Household Income in US$')
```



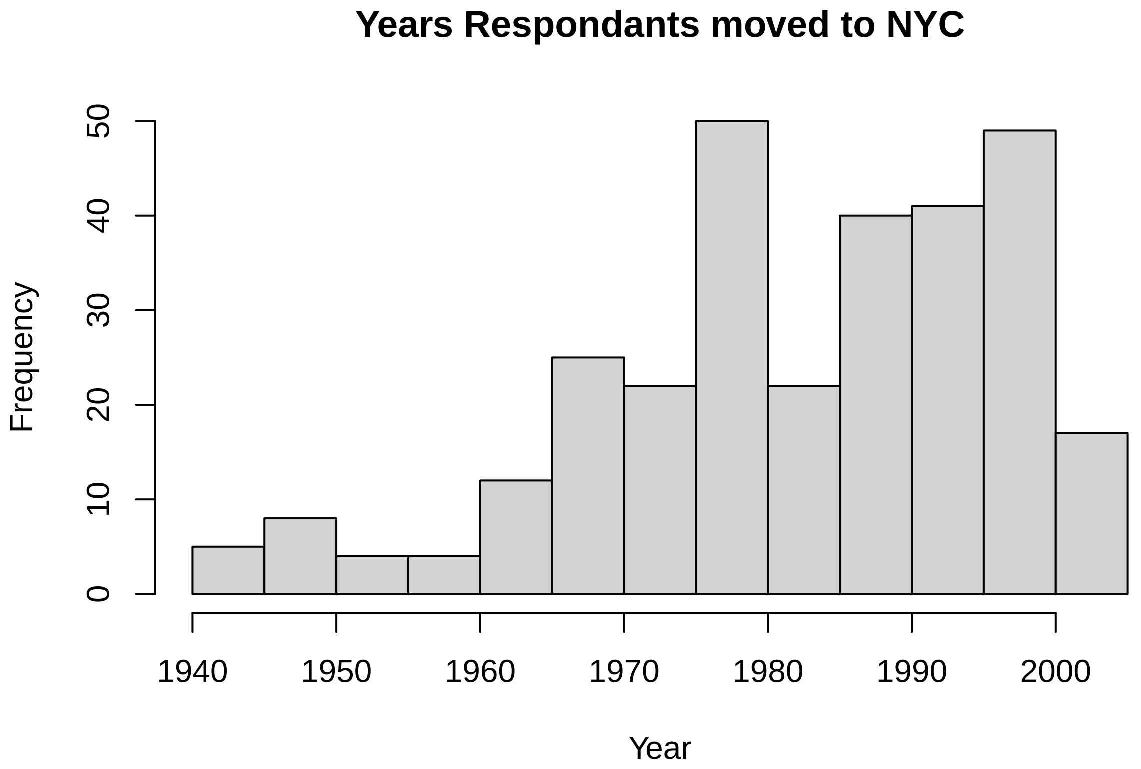
```
hist(nyc$Age,
     main= 'Ages of Respondants of Survey',
     xlab='Age in Years')
```



```
hist(nyc$MaintenanceDef,
     main= 'Frequency of Maintenance Deficiencies',
     xlab='No. Of Maintenance Deficiencies')
```



```r
hist(nyc$NYCMove,
     main= 'Years Respondants moved to NYC',
     xlab='Year')
```



We supplement the univariate graphical summaries with numerical summaries as follows:

For **Income:**

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

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

```
sd(nyc$Income)
```

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

Based on the Histogram and the numerical summaries we make the following obervations about **Income**

For the variable "Income" our graphical representation, presented as a histogram, predominantly indicates a unimodal distribution with a noticeable right skew. In a numerical summary of the data, we observe incomes in New York City ranging from a minimum of $1,440 to a peak of $98,000, with a central tendency captured by a median of $39,000.

Given the diverse spectrum of professions in New York City, such a distribution and the resultant standard deviation of $24,204.1 are both plausible and expected. Additionally, the income distribution doesn't exhibit any apparent outliers.

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
```

```
sd(nyc$Age)
```

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

For the variable "Age" of the respondents, the histogram showcases a predominantly unimodal distribution with a slight rightward skew. Such a distribution is consistent with the attraction of the younger population to the city, given its myriad of opportunities, hence rationalizing the right skew.

The numerical summary reveals a diverse yet plausible age range spanning from 26 to 85 years. The central age, represented by the median, stands at 49 years, and the standard deviation being 12.43678 years. Upon close examination, the age distribution seems to be devoid of any significant outliers.

For **MaintenanceDef:**

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

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

```
sd(nyc$MaintenanceDef)
```

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

For the variable 'MaintenanceDef,' representing the number of maintenance deficiencies in residences between 2002 and 2005, the histogram shows a pronounced rightward skew which is favorable from a respondent's viewpoint, indicating fewer deficiencies.

Numerically, the data spans a range from a minimum of 0 deficiencies to a maximum of 8, with a standard deviation of 1.62. Upon thorough analysis, there seems to be an absence of significant outliers. Next, we will turn our attention to the year respondents relocated to NYC.

For **NYCMove:**

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

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

```
sd(nyc$NYCMove)
```

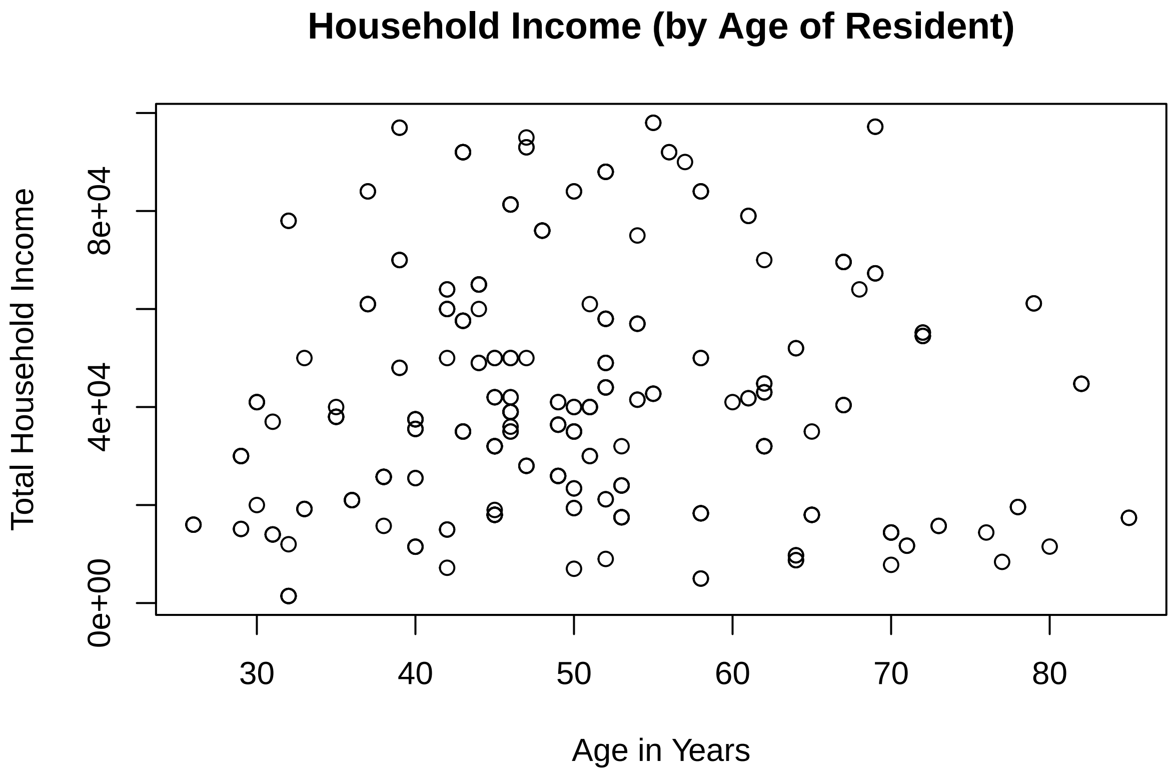
```
## [1] 14.13746
```

For the variable “NYCMove”, representing the year respondents relocated to New York City, the histogram showcases a distinct unimodal distribution with a leftward skew. This skew is justifiable given the increased influx of people into NYC in more recent years. The data indicates that the earliest a respondent moved to the city was in 1942, while the most recent move was recorded in 2004. The dispersion in the years is captured by a standard deviation of approximately fourteen years. A detailed inspection reveals no significant outliers in the distribution.

Bivariate Exploration

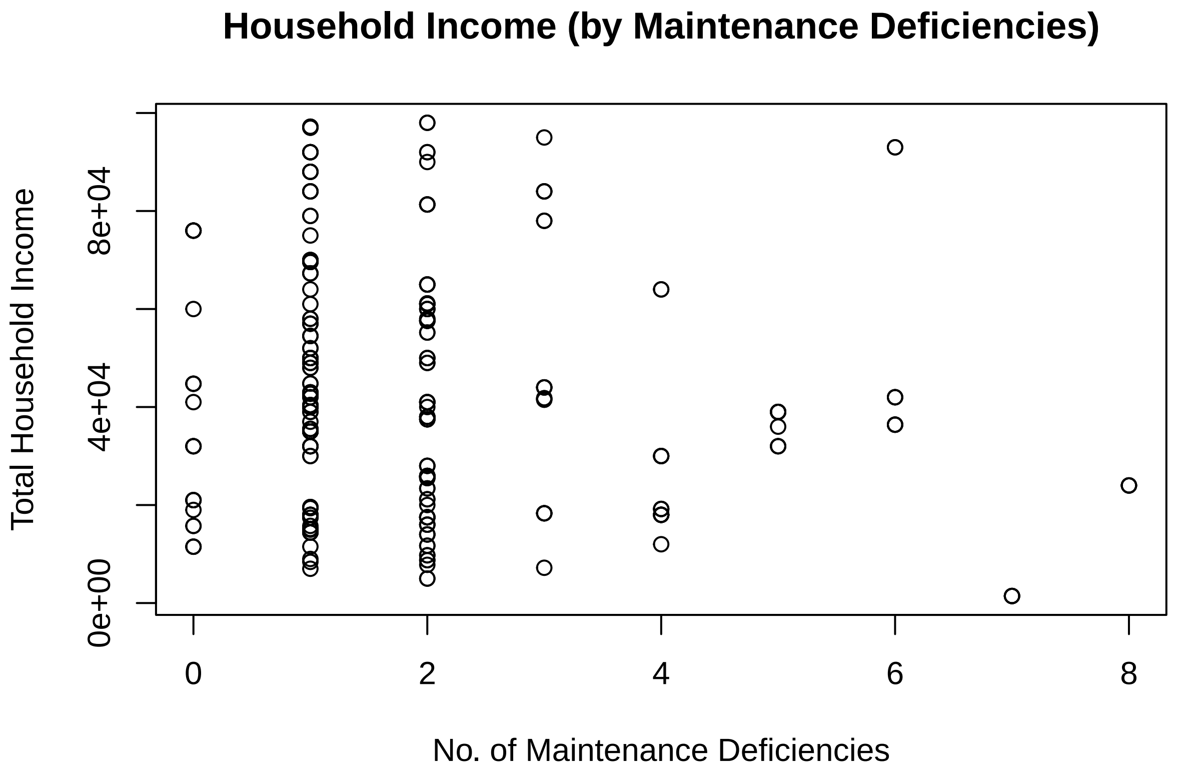
Now that we understand the distribution of the individual variables in this data, we can graphically plot and understand how each predictor is associated with the response **Income**, as follows:

```
plot(nyc$Income ~ nyc$Age,
     data=nyc,
     main ="Household Income (by Age of Resident)",
     xlab ="Age in Years",
     ylab ="Total Household Income")
```



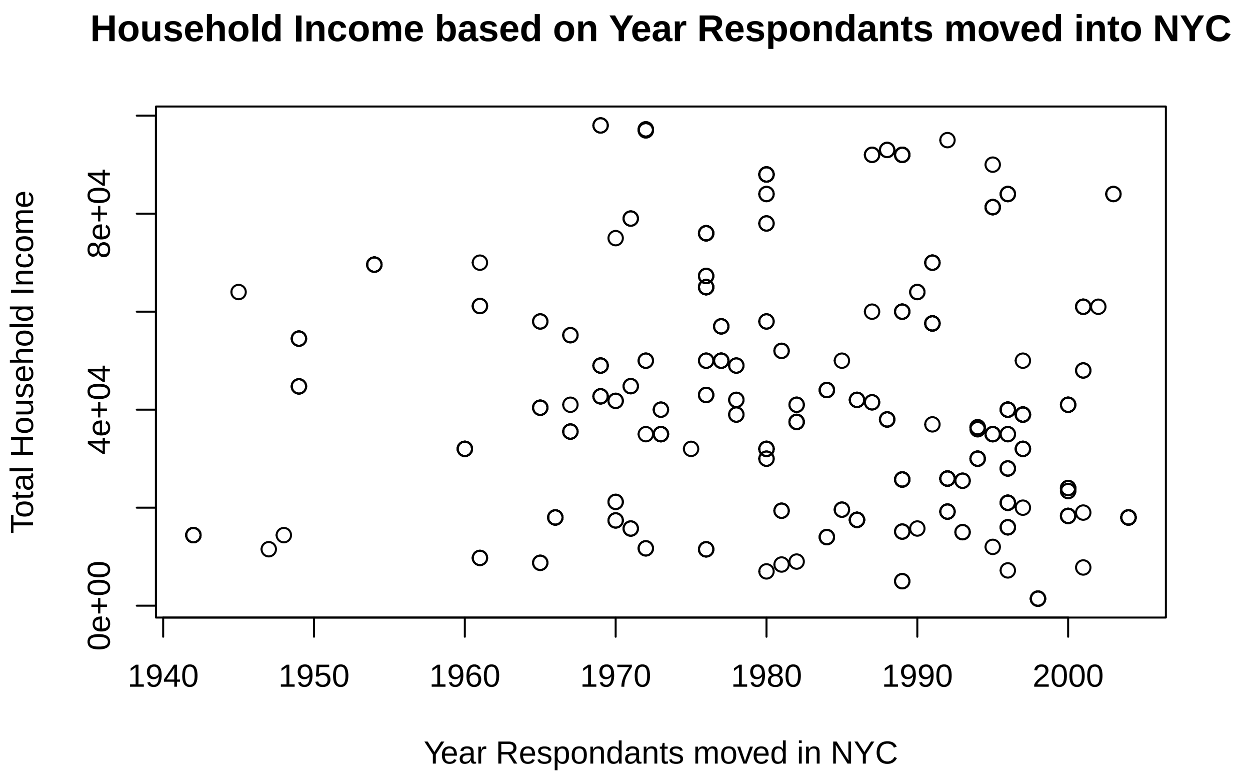
The scatter plot depicted illustrates data points extensively dispersed across the graph, suggesting a non-linear relationship between age and income. From the visual representation, it’s evident that there isn’t a discernible correlation, especially not a linear one, between the two variables: Age and Income.

```
plot(nyc$Income ~ nyc$MaintenanceDef,
     data=nyc,
     main ="Household Income (by Maintenance Deficiencies)",
     xlab ="No. of Maintenance Deficiencies",
     ylab ="Total Household Income")
```



The scatter plot displayed conveys data points that are largely clustered in the lower range of maintenance deficiencies. This suggests that households with fewer maintenance deficiencies tend to have varying incomes, but as the number of deficiencies increases, the income data becomes sparser. There isn't a clear linear relationship between maintenance deficiencies and household income, indicating an absence of a direct correlation between the two variables.

```
plot(nyc$Income ~ nyc$NYCMove,
     data=nyc,
     main ="Household Income based on Year Respondants moved into NYC",
     xlab ="Year Respondants moved in NYC",
     ylab ="Total Household Income")
```



The scatter plot presented depicts the relationship between the year respondents moved into NYC and their household income. The data points are spread out across the years, with a noticeable density in the 1980s and 1990s. While the incomes seem varied for each decade, there's a slight increase in higher incomes for those moving to NYC in the 1990s and early 2000s. However, there isn't a consistent linear trend, suggesting that the year of moving to NYC might not be a strong correlation with the household income.

Modeling

After exploring and visualizing the relationships among our variables, we now turn to building a linear regression model to predict Household Income. We start by looking at the histogram of our response variable, which as observed before has a noticeable right skew indicating that a transformation might be needed.

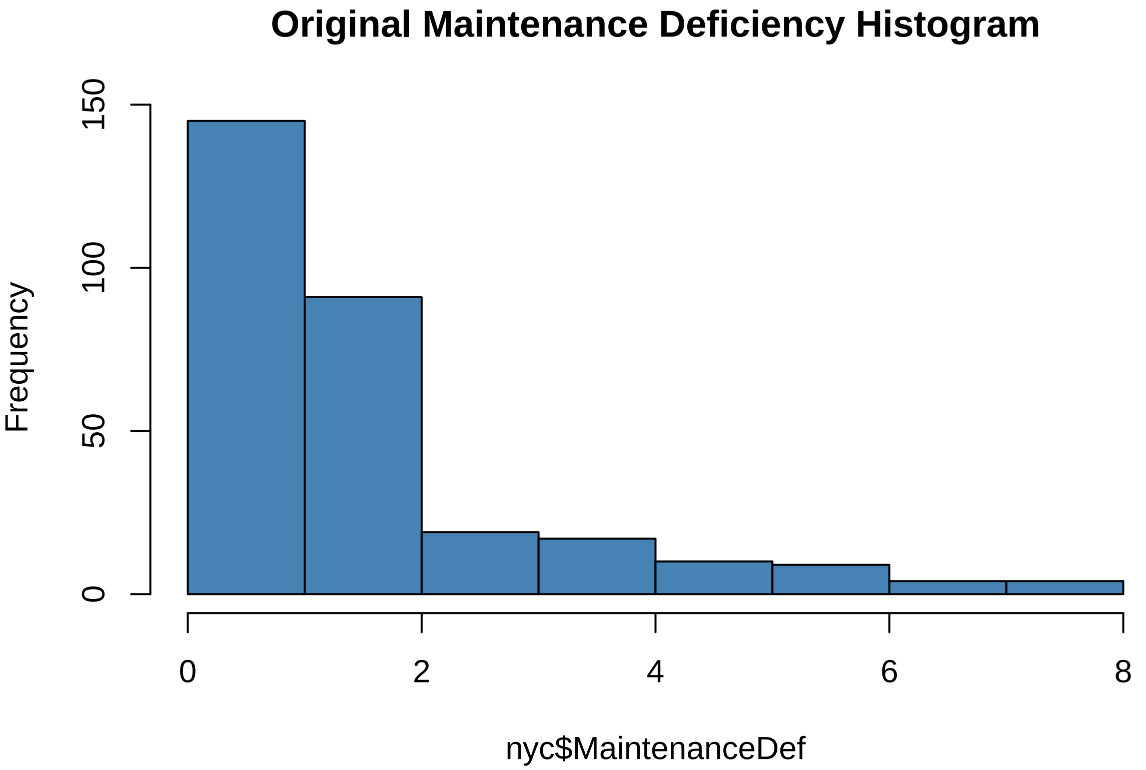
For a start we know that our linear regression equation should in the end be of the format,

```
beta0 + beta1(Age) + beta2(MaintenanceDef) + beta3(NYCMove) + E
```

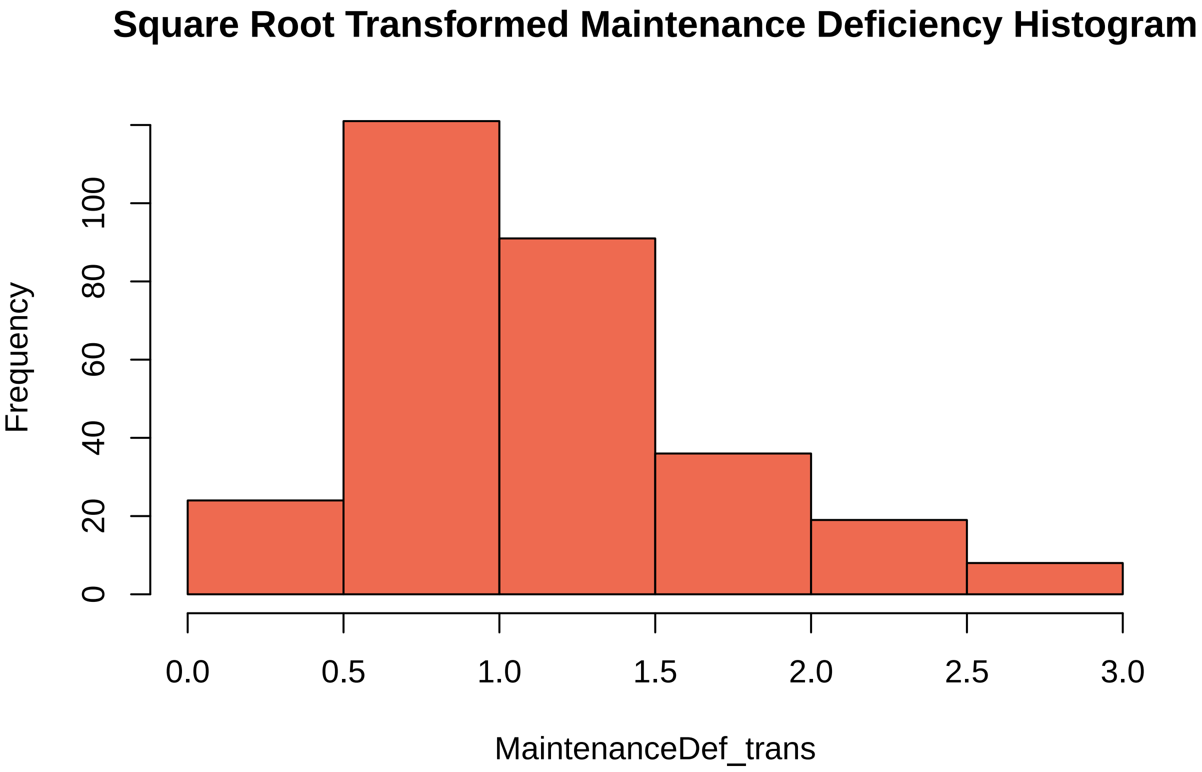
In order to have a rough idea we will see the level of correlations between the explanatory variables and the response variable,

We will begin with performing a square root transformation on MaintenanceDef, owing to its high skewedness compared to others as well as its strong non-lineraity when plotted against Income

```r
MaintenanceDef_trans <- sqrt(nyc$MaintenanceDef)
hist(nyc$MaintenanceDef, col='steelblue', main='Original Maintenance Deficiency Histogram')
```



```
hist(MaintenanceDef_trans, col='coral2', main='Square Root Transformed Maintenance Deficiency Histogram
```



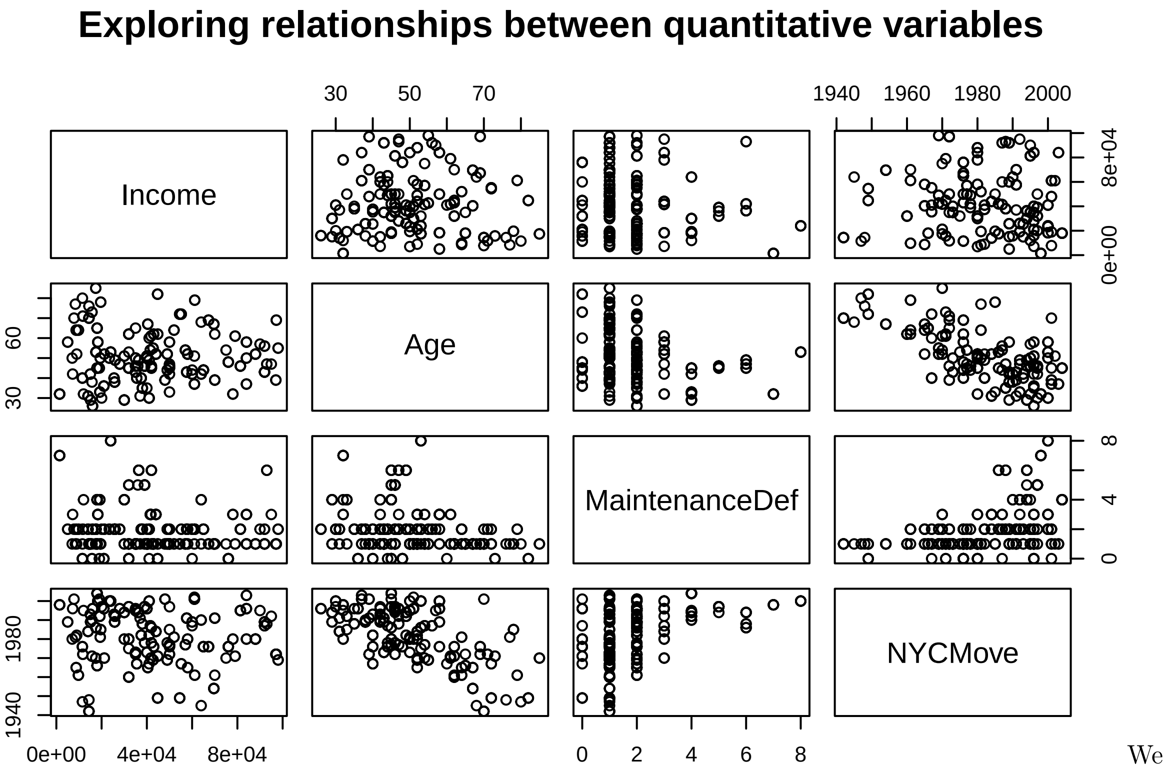
The trasnformed MaintenanceDef Histogram now looks much more normal as compared to the original one. So we can create the linear model, whose summary is as follows,

We also need to check for risk of multicollinearlity since there are 3 explanatory variables and there is always a high chance for it. We can do this by plotting the variables side by side as well as looking at their correlation matrix to have a generic idea.

```
cor(dplyr::select(nyc, Income, Age, MaintenanceDef, NYCMove))
```

```
##                     Income         Age MaintenanceDef    NYCMove
## Income          1.00000000  0.03593162     -0.1681017 -0.1009987
## Age             0.03593162  1.00000000     -0.2486687 -0.6365920
## MaintenanceDef -0.16810175 -0.24866870      1.0000000  0.4563387
## NYCMove        -0.10099865 -0.63659204      0.4563387  1.0000000
```

```
pairs(nyc,
      main='Exploring relationships between quantitative variables')
```



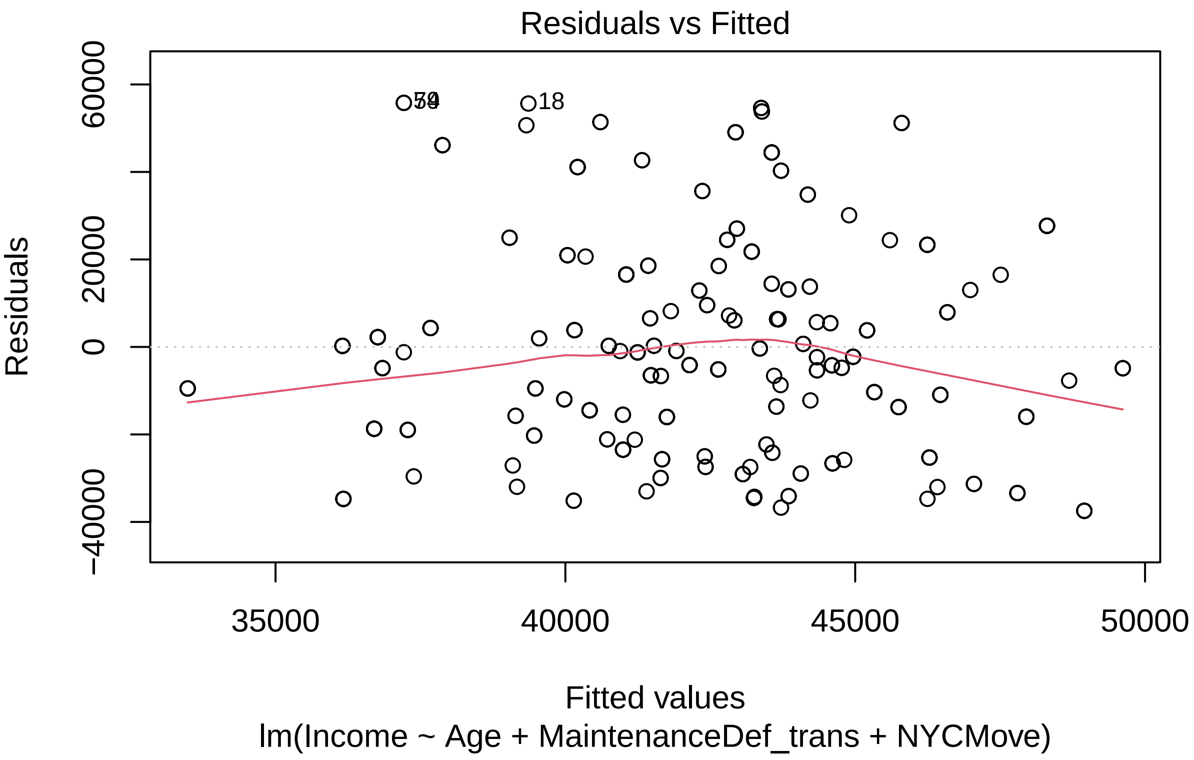
We can observe there seems to be a strong linear corrleation between “Age” and “NYCMove” alerting us of a potential multicollinearlity, but to formally confirm we will have a look at the VIF values for all the variables.

```
model_trans <-lm(Income~Age+MaintenanceDef_trans+NYCMove,
                 data =nyc)
vif(model_trans)
```

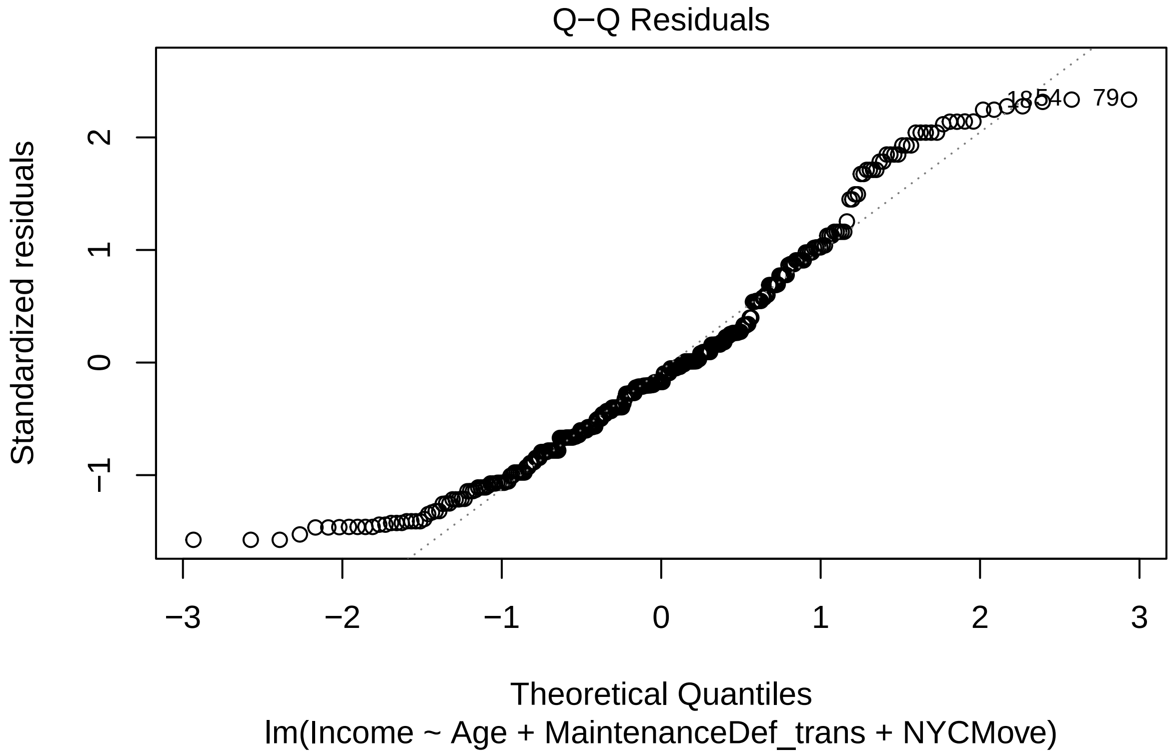
```
##                  Age MaintenanceDef_trans              NYCMove
##             1.687276             1.233234             1.960362
```

Given that the VIF values for all variables are below 2.5, we can confidently eliminate the possibility of severe multicollinearity. This also indicates that the transformed “MaintenanceDef” data is apt for inclusion in our analysis. Next, we’ll examine the residual diagnostic plots for all the predictor variables in our updated model.

```
plot(model_trans,
     which=1)
```



```r
plot(model_trans,
     which=2)
```



While there are deviations between 1 and 2 and a few outliers, the qq plot largely supports the assumption of normality, as the majority of points align closely with the line. The residual plot does exhibit notable outliers on either side, yet it maintains a consistent spread throughout. This graph also underscores the assumption of mean zero, given that the residuals scatter evenly both above and beneath the zero line, showing no discernible trend. The lack of a clear pattern among the residuals confirms their independence. The transformation of the maintenance variable proved beneficial, ensuring that all required residual assumptions were met. Following is a summary of our regression analysis that represents the finalized model.

Thus we can write the final model and its associated summary is as follows,

```
MaintenanceDef <- MaintenanceDef_trans
finalModel <-lm(Income~Age+MaintenanceDef+NYCMove,
                data =nyc)
summary(finalModel)
```

```
##
## 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
```

The Regression model thus becomes,

```
Income = 237408.41 - 71.98(Age) - 2273.22(MaintenanceDef) -94.34(NYCMove)
```

Only one coefficient, which pertains to Maintenance Deficiencies raised to 1/2, is statistically significant with a p-value below the 5% alpha level.

On average, the coefficient for the transformed Maintenance variable suggests that for every decrease in household income by $2,273.2, there is a unit increase in Maintenance Deficiencies.

The significance of this final model is underscored by its F-test p-value of 0.03005, which is beneath the 5% alpha threshold.

Furthermore, the robustness of the data analysis is bolstered by the fact that all three explanatory variables showcase negative coefficients and exhibit minimal collinearity, as evidenced by their VIF values being below 2.5[We also remark that these negative values are also confirmed from our EDA]

Prediction

We can perdict the household income of a 53 year old respondent who moved to NYC in 1987 and has had 3 maintenance deficiencies using the following equation:

```
Income = 237408.41 - 2273.22(MaintenanceDef) -94.34(NYCMove)
```

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

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

[We note that the prior prediction could also be obtained with the predict() function.]

The predicted Household Income of a 53 year old respondant who moved to NYC in 1987 and has had 3 maintenance deficiences to be $39,320. [We remark that this value is considererd as a low household income as it is lower than the average Mean Household Income by almost half standard deviation for the dataset]

Discussion

In our thorough exploration, after adjusting the 'MaintenanceDef' variable, it became evident that Household Income in New York City (our target variable) is intricately connected to Maintenance Deficiencies (our predictor). Notably, we didn't observe any serious multicollinearity issues, affirming Maintenance Deficiencies as the primary influential factor in our refined regression model.

Housing research, particularly in metropolises like New York City, is of paramount importance. Here, housing often seems to be a luxury reserved for the exceedingly affluent. Our dataset offers meaningful insights into this dynamic, but it's not without limitations. A notable gap is the absence of external contextual elements that could potentially reshape our understanding of the predictors. This limitation becomes especially glaring when considering that variables such as 'Year the Respondent Moved to NYC' and 'Age' don't significantly correlate with 'Income'.

Reflecting on our study, it's clear that the economic dynamics of housing in New York City are influenced by various factors, with Maintenance Deficiencies playing a pivotal role. However, the role of other variables like 'Age' and 'Year the Respondent Moved to NYC' remains ambiguous due to the dataset's limitations. Going forward, a more comprehensive dataset that includes other influencing factors might offer a richer understanding.

In the grand scheme, studies like ours emphasize the importance of continuous research in urban housing dynamics. As the city evolves, so do its challenges, and a deeper understanding will benefit both policymakers and residents.