Assessing Predictors of Income in NYC Households

Contents


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

Introduction

The motivation for this topic is to study New York City household income, which is important because understanding urban household income and factors that relate to income, like age, can provide importance to policymakers who aim to create fair policies for people in New York City (such as creating regulations and adjusting property taxes). The impact of this project on the world is that New York City is the largest city in the United States, so learning how New Yorker's income are related to other factors like age and the year moved to NYC is important.

The client of this study also wants to predict the income of a household with three maintenance deficiencies, whose respondent's age is 53 and who moved to NYC in 1987.

Reference: The New York City Housing and Vacancy Survey, https://www.census.gov/programs-surveys/nychvs.html

Exploratory Data Analysis

Data Set Description To begin the exploratory data analysis of the New York City data, we will first discuss the background information on how the data was collected. The data comes from The New York City Housing and Vacancy Survey, which is conducted every three years. The survey is done to learn about New

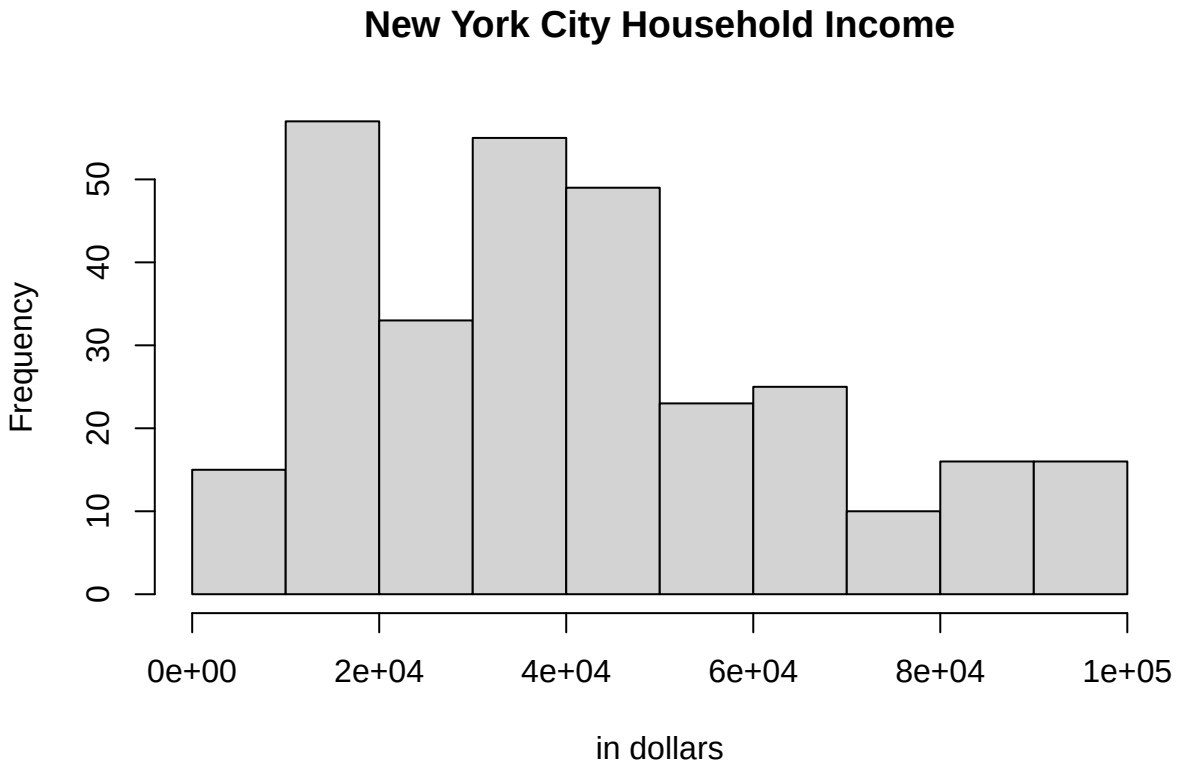
Yorker's housing conditions. The survey also has a high response rate and the current project uses a sample of some of the survey data.

Now, we will discuss what a singular observation in the data would look like. In each single household observation, there are four variables provided in the sample through the survey, these being total household income in dollars (shows as Income in the data), respondent age in years (shows as Age in the data), the year the respondent moved to New York City (shows as NYCMove in the data), and maintenance deficiencies, known as the number of maintenance deficiencies of the residence, between 2002 and 2005 (shows as MaintenanceDef in the data).

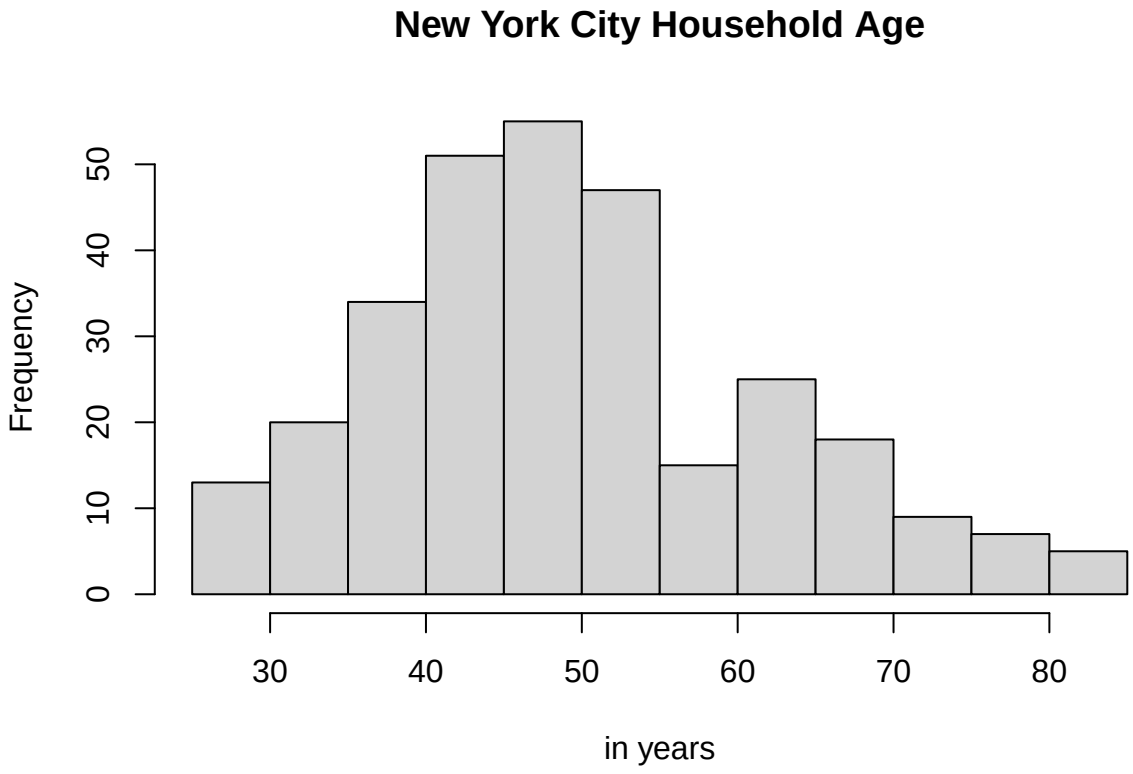
Use written text along with numerical summaries and graphs. Perform univariate and bivariate EDA to summarize/explore the variables and their relationships.

Univariabe Exploratory Data Analysis Now, we look to univariate EDA for each of the variables. Note that all the variables being referrenced in this study are quantitative.

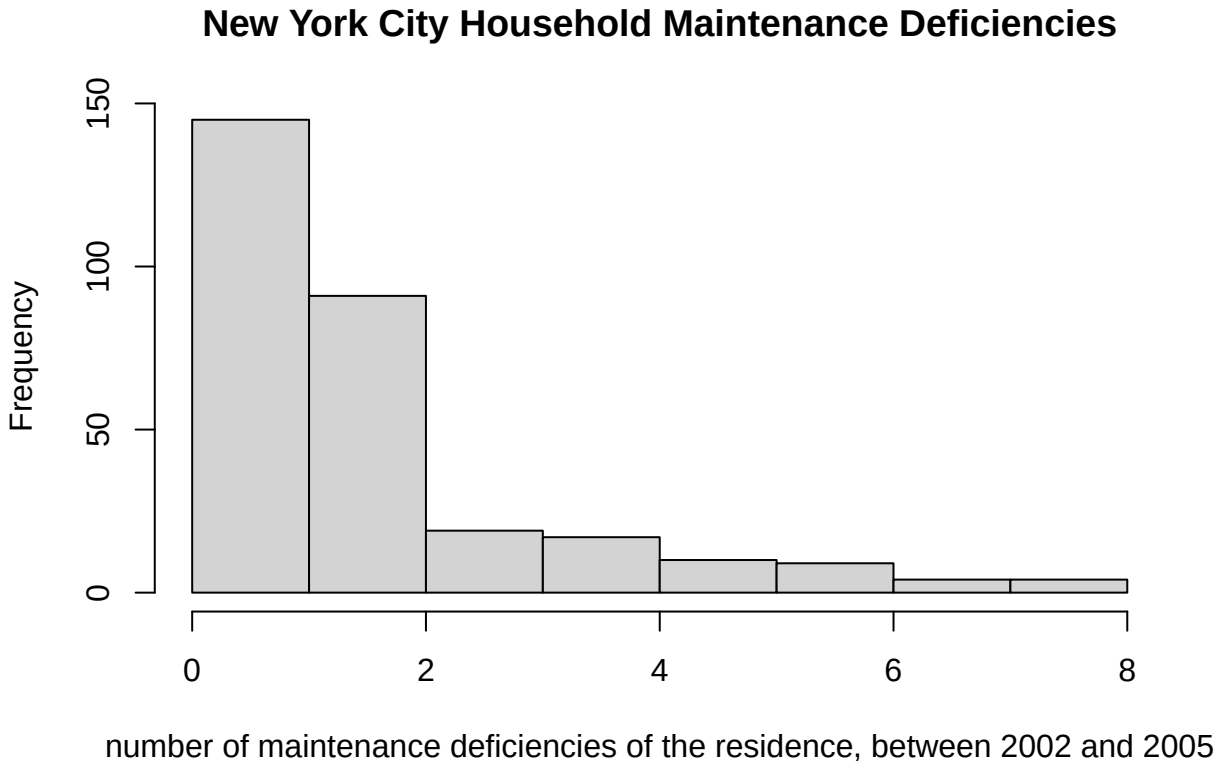
```
hist(nyc$Income,
     main = "New York City Household Income",
     xlab =  "in dollars")
```



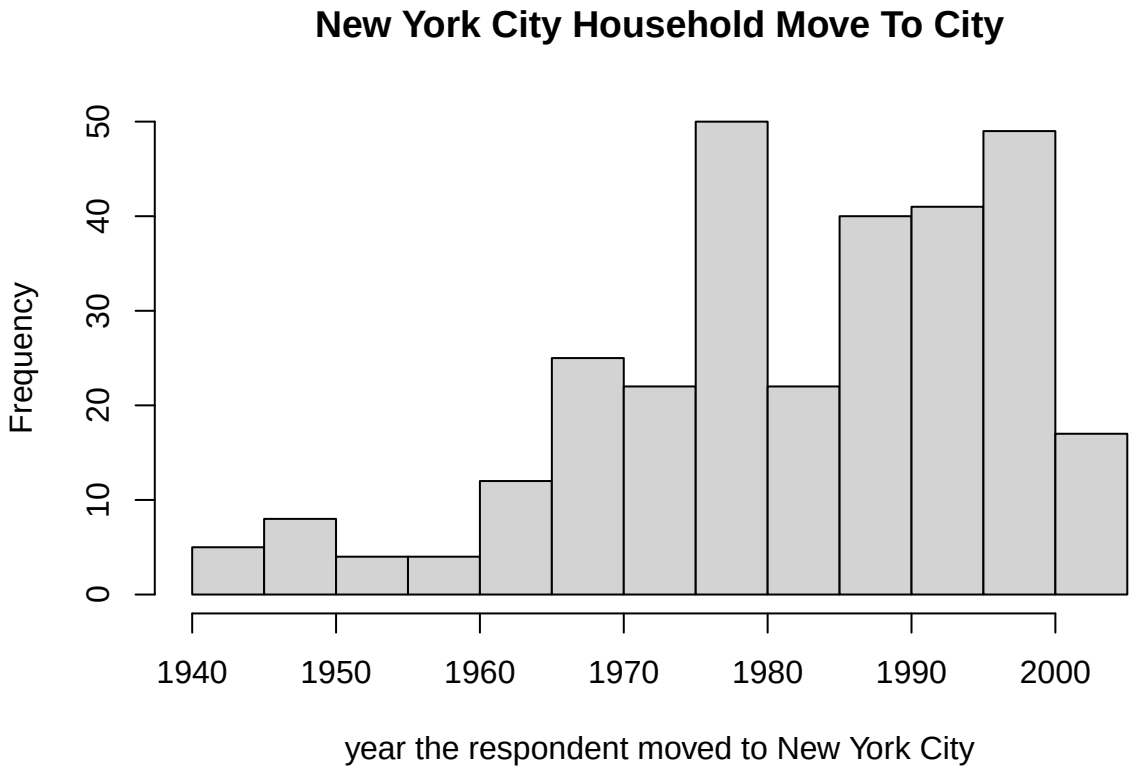

```
hist(nyc$Age,
     main = "New York City Household Age",
     xlab =  "in years")
```




```
hist(nyc$MaintenanceDef,
     main = "New York City Household Maintenance Deficiencies",
     xlab =  "number of maintenance deficiencies of the residence, between 2002 and 2005")
```



```
hist(nyc$NYCMove,
     main = "New York City Household Move To City",
     xlab =  "year the respondent moved to New York City")
```



Additionally, we provide summaries of each of the variable data sets, as follows.

Summary of Income

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

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

Summary of Age

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

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

Summary of Maintenance Deficiencies

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

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

Summary of NYCMove

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

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

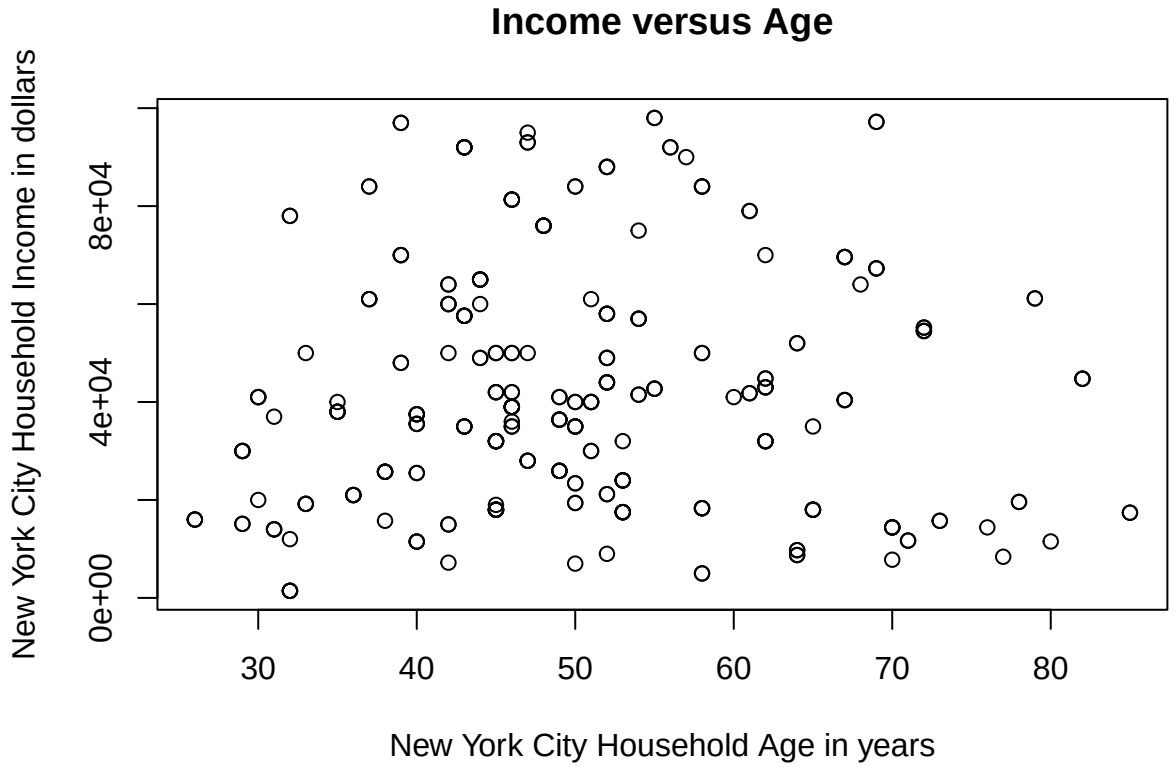
We can see that income variable has a skewed right distribution and is unimodal, the age variable has a

slightly symmetric skew and is unimodal, the maintenance deficiencies variable has a strong right skew and is unimodal, and the NYC move variable has a left skew distribution. It is difficult to tell if the NYC Move variable is unimodal or bi-modal considering the frequency of data of people moving to NYC in the end-half of the 1970s. Neverthe less, the graph appears bimodal, however we would need some more data to make a better characterization.

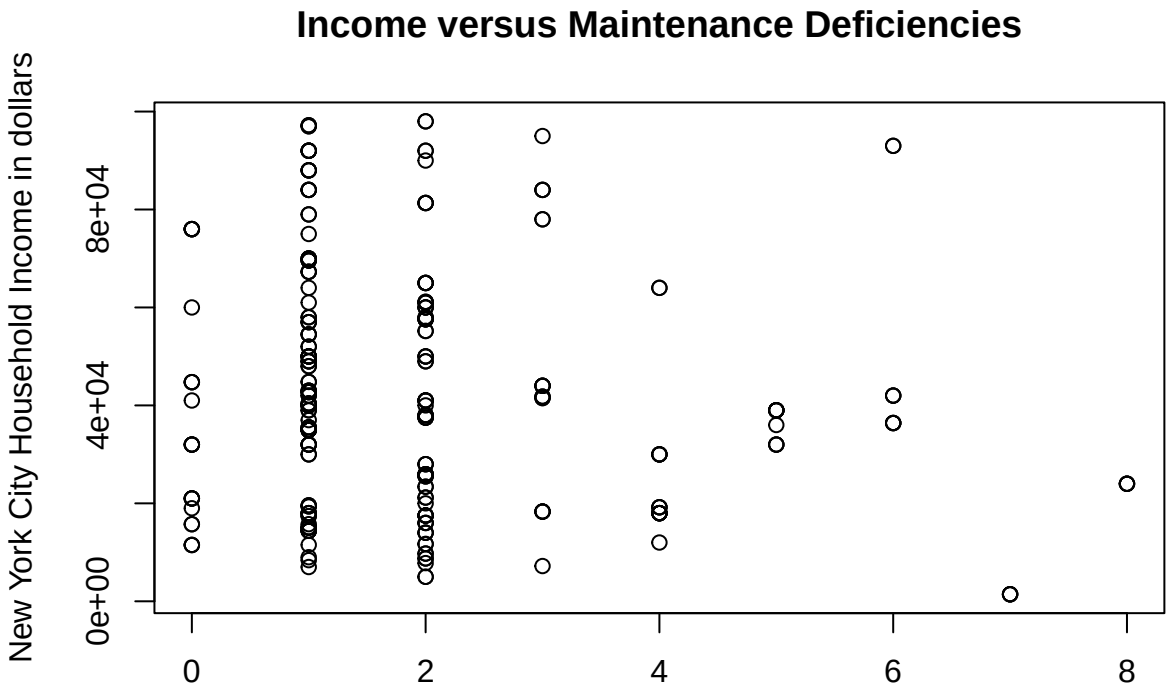
The mean and median income are 42266 and 39000, respectively. The mean and median age are 50.03 and 49, respectively. The mean and median income maintenance deficiencies are 2 and 1.98, respectively. The mean and median NYC move year are 1983 and 1985.

Bivariate Exploratory Data Analysis Now, we will discuss bivariate EDA, comparing income with age, maintenance deficiencies, and NYC move time. Since all variables are quantitative, we will use scatterplots for each comparison. Income is the response variable.

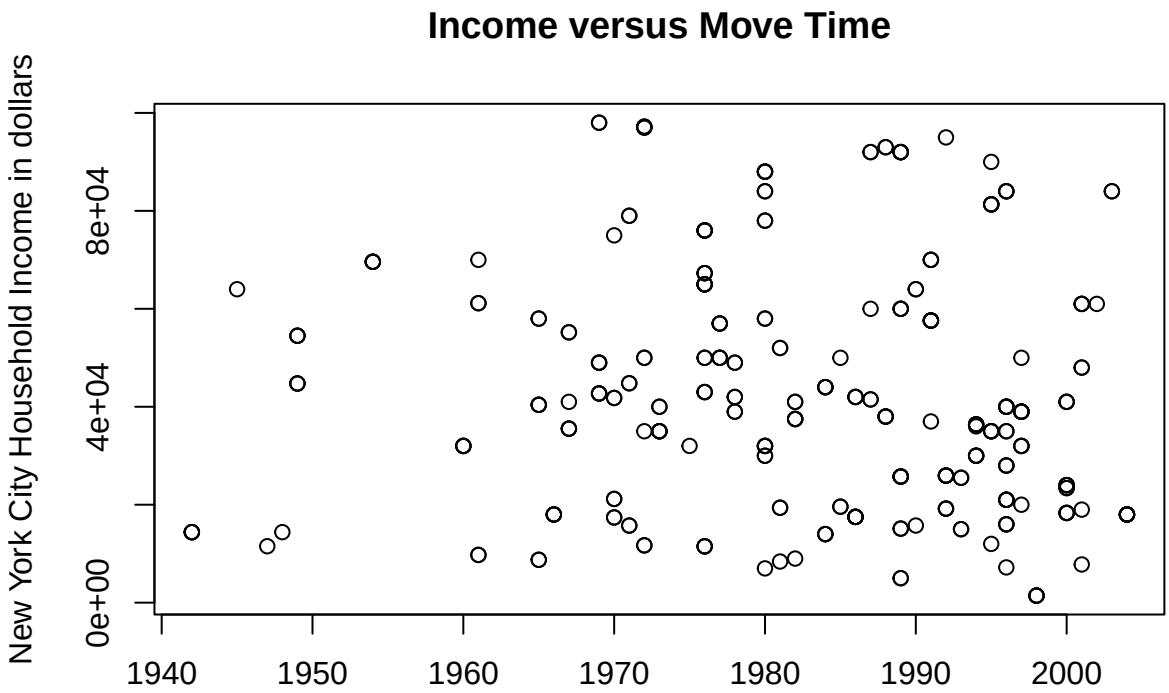
```
plot(Income ~ Age,
     data = nyc,
     main = "Income versus Age",
     xlab = "New York City Household Age in years",
     ylab = "New York City Household Income in dollars")
```




```
plot(Income ~ MaintenanceDef,
     data = nyc,
     main = "Income versus Maintenance Deficiencies",
     xlab = "Number of maintenance deficiencies of the residence, between 2002 and 2005",
     ylab = "New York City Household Income in dollars")
```




```
plot(Income ~ NYCMove,
     data = nyc,
     main = "Income versus Move Time",
     xlab = "Year the respondent moved to New York City",
     ylab = "New York City Household Income in dollars")
```



By analyzing each scatterplot, it is apparent that age is positively related with income. That is, as age increases, household income also rises. This relationship between age and income is not very strong, however. Additionally, it seems that the relationship between income and move time (NYCMove) is unclear, if there is even a relation. Third, the association between income and maintenance deficiencies is also unclear. That is, we do not know if the association is positive, negative, or not significantly related.

Modeling

Now that we have conducted exploratory data analysis, we build a linear regression model that can predict a household's income.

To start, we will use the correlation matrix since it is part of the exploratory data analysis for multiple regression, and all the variables are quantitative. The correlation coefficients below are shown, relating the response variable (Income) to the explanatory variables, and relating the explanatory variables to themselves too.

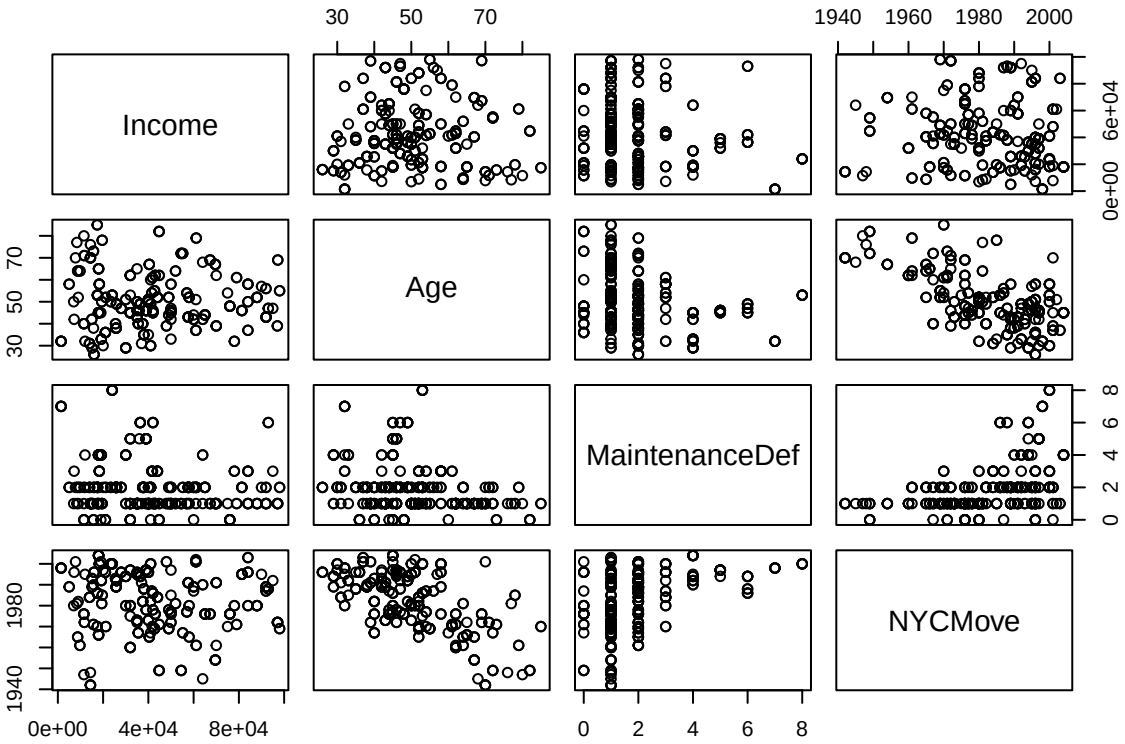
```
round(cor(nyc),
    digits = 2)
```

```
##                Income   Age MaintenanceDef NYCMove
## Income           1.00  0.04          -0.17   -0.10
## Age              0.04  1.00          -0.25   -0.64
## MaintenanceDef  -0.17 -0.25           1.00    0.46
## NYCMove         -0.10 -0.64           0.46    1.00
```

Here, we see the correlation matrix. It appears that Income is only slightly negatively correlated with the maintenance and NYCMove variables, and the correlation between Income and Age is 0.04. We also note that Age and NYCMove are negatively correlated with each other at a value of -0.64. Further, Maintenance and NYCMove are positively related with each other at a value of 0.46. Of the predictors, MaintenanceDef has the strongest relationship with Income.

To support the correlation matrix, we also use a pairs plot, as follows.

```
pairs(nyc)
```



In the pairs plot, we notice the negative relation between Age and NYCMove (a rise in NYCMove year results in a decrease in age) and the relation between NYCMove and MaintenanceDef (a rise in NYCMove year results in more maintenance deficiencies). Given these associations, there may be an issue of multicollinearity.

We want to not have high multicollinearity because that would result in two explanatory variables being highly related with one another, potentially resulting in incorrect results. So, we will use the variance inflation factor (vif) for each explanatory variable.

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

car::vif(income.full.mod)
```

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

Since none of the vif values are above 2.5, there are no significant issues with high multicollinearity. Therefore, we continue building the multiple linear regression model with all variables.

Below, we have a multiple linear regression model predicting income from Age, MaintenanceDef, and NYCMove.

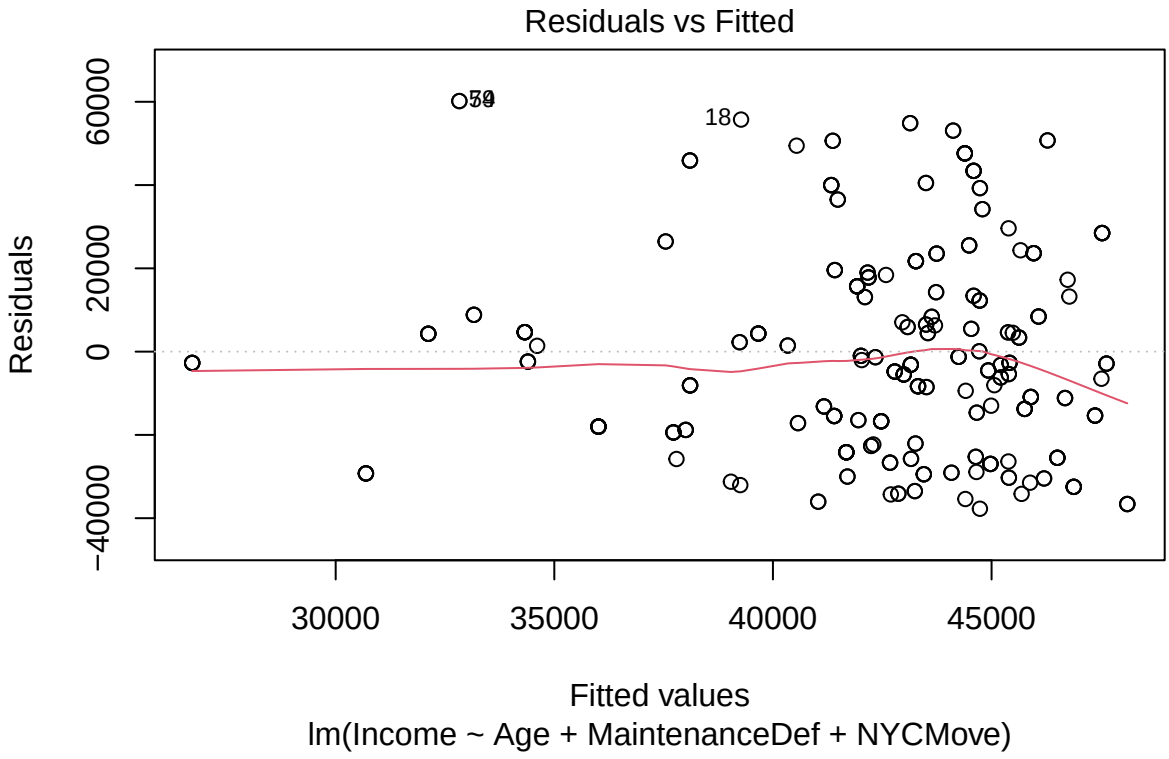
```
nyc$log.Income <- log(nyc$Income)

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

summary(income.full.mod)
```

```
##
## Call:
```

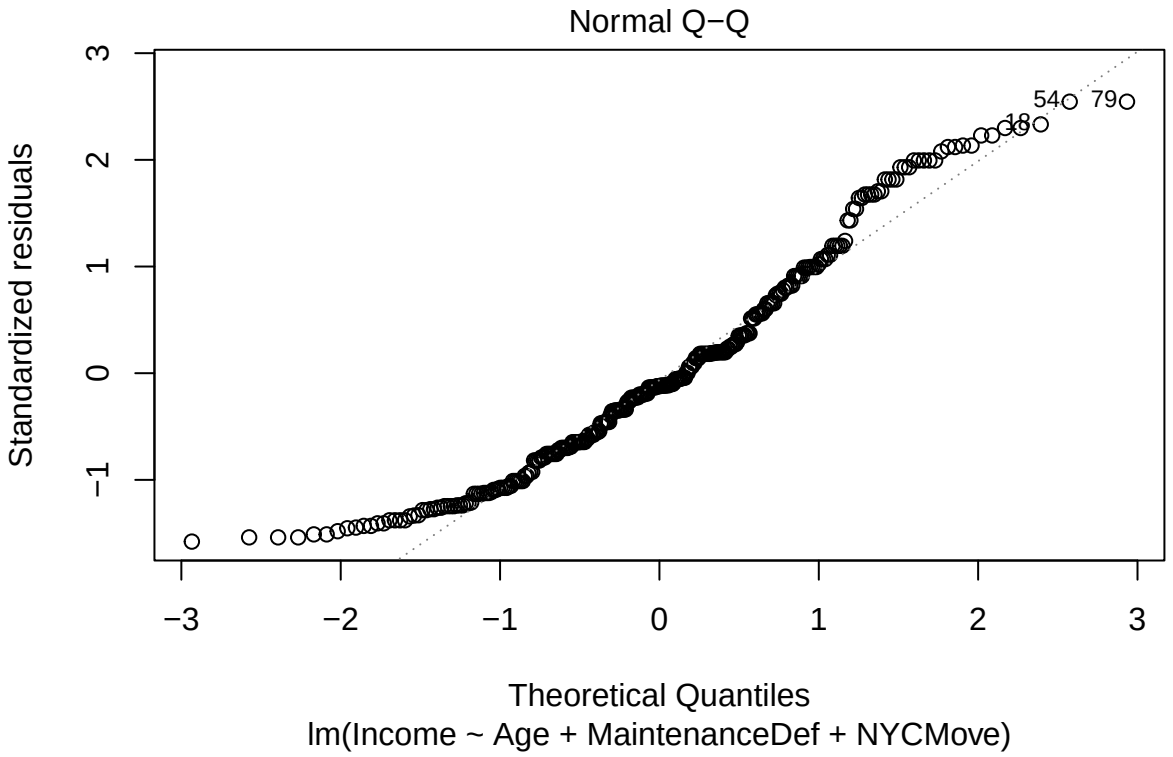
```
## lm(formula = Income ~ Age + MaintenanceDef + NYCMove, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37734 -18010  -2878  14971  60171
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    237408.41  278939.01   0.851   0.3954
## 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
```

The adjusted R-squared is quite low, and unfortunately from testing other linear models, there does not seem an exist an easy way to increase the R-squared value. The reason for the low R-squared value may be because income is dependent on many factors outside of just age, maintenence deficiencies, and moving year to New York. From the correlation matrix, all of the variables show some degree of relation, so each is included in the multiple linear regression model above. The F-statistics of the model yields a p-value that is 0.03005, which is significant and sufficient for the model.

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




```
plot(income.full.mod,
     which = 2)
```



Above, we have the Residual and Normal Q-Q Plots. It appears that the mean zero assumption is valid since the residuals on the plot look reasonably centered near the zero line. The independence assumption seems valid as the residuals are patternless from up to down. The normality assumption is valid, since the Q-Q Plot points are close to the line. The equal spread assumption look sufficient for most of the data. There is greater spread in the residuals on the left side, however transformation used, such as logorithms, made the spread relatively worse, so we used no transformation. Specifically on the linearity assumption, there appears to be no curve or clear trend.

Attempts to transform the data using logorithms resulted in worse diagnostics.

Prediction

We want to predict the income of a person who reports a household with three maintenance deficiencies with an age of 53 and a move year of 1987. Following the multiple linear regression model above, we have:

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

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

The value of a person who is 53 with 3 maintenance deficiencies and a move year of 1987 is **39320.23 dollars**.

Discussion

The overall conclusions are that age, maintenance deficiencies, and NYC move year all contribute toward the prediction of income. Transformation attempts resulted in worse diagnostics, so a multiple linear regression model based on all original quantitative variables was used. In critiquing the analysis, the limitation included the low R-squared value. This limitation may be because income is predicted by a variety of factors outside of the variables used in this study. Issues with the data included a relatively small sample size. In comparison to the full data set, the selected data was a far smaller percent of what was actually available. Diagnostics issues included the residual plot. Further study will need to be done on making a model with strong equal

spread on the Residual Plot. Future projects can incorporate a more detailed analysis of more variables that contribute to income in New York City, as income is dependent on many different factors. Data analysts could thus build on this work by sourcing other variables.