Predicting Respondent Income from NYC Housing Conditions and Demographic Features

Raaka Mukhopadhyay
rmukhopa

Due Wed, Oct 21, at 8:00PM (Pittburgh time)

Contents


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

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

Introduction

At the end of the 19th century, many individuals that lived in New York City resided in tenements, as shown in the photojournalism study "How the Other Half Lives." Now, more than 100 years later, New York City is home to over 8 million individuals. In order to better comprehend the current housing conditions, we will look at the relationship between respondent income and several predictors. As a result, in this paper, we will aim to have a better understanding of the current living conditions of New Yorkers in "the other half.""

https://www.census.gov/quickfacts/newyorkcitynewyork?

Exploratory Data Analysis

Description of Data

The New York City Housing and Vacancy Survey is performed every three years; the survey is known to be designed thoughtfully and has a high response rate. We have been given a sample of data from a

borough/sub-borough pair. The sample contains 299 responses and four variables: Income, Age, Maintenance, and Year Moved to NYC. Since we are interested in respondent income, we will take a closer look at income and its relationship with three explanatory variables: Age, Maintenance, and Year Moved to NYC. The following presents a description of each of the variables:

`Income`: total household income of respondent (measured in dollars) `Age`: age of respondent (measured in years) `MaintenanceDef`: respondent's total maintenance deficiencies between 2002 and 2005 `NYCMove`: the year the respondent moved to New York City (measured in discretely in years)

To provide a better understanding of the data, we show the first few lines below:

```
head(nyc)
```

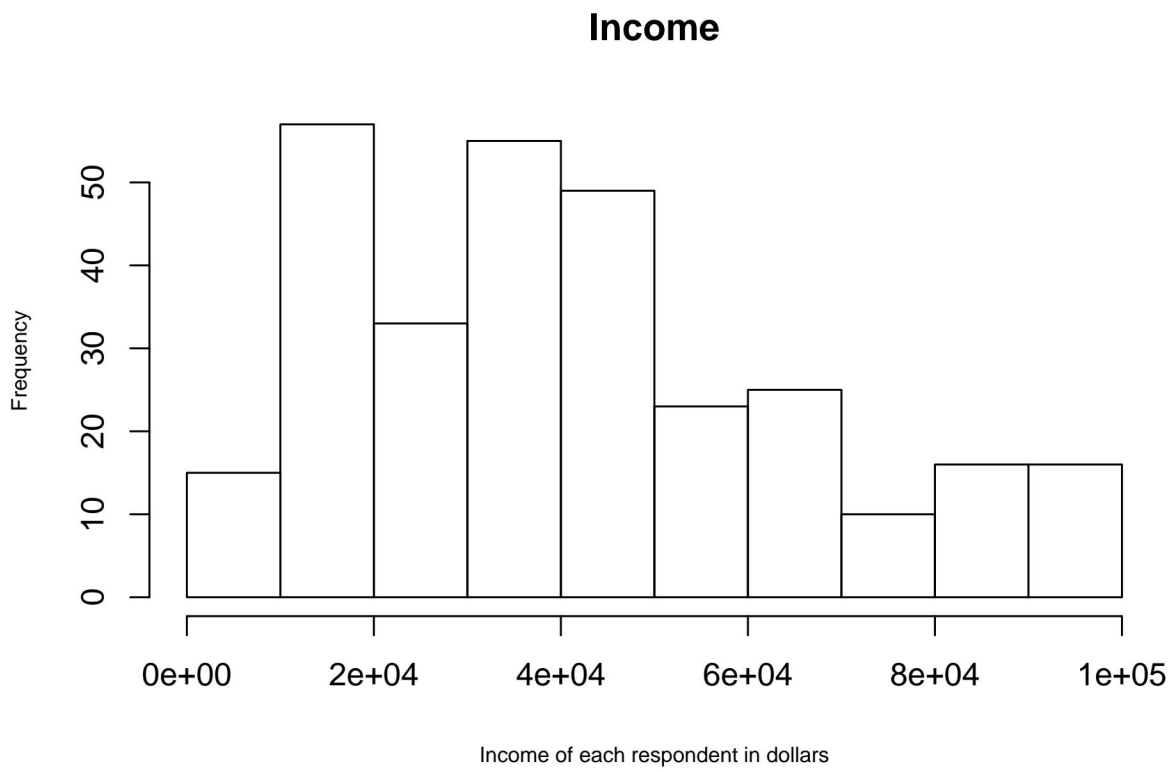
```
## # A tibble: 6 x 4
##   Income   Age MaintenanceDef NYCMove
##    <dbl> <dbl>          <dbl>   <dbl>
## 1   8400    77              1    1981
## 2  17510    53              2    1986
## 3  19200    33              4    1992
## 4  42717    55              1    1969
## 5   5000    58            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

Univariate EDA

In order to better understand both the predictor variables and the response variable, we will perform a univariate data analysis. Since all the variables are quantitative, we will use a histogram to learn more about each variable.

EDA on Y (Income)

```
hist(nyc$Income,
  main = "Income",
  xlab = "Income of each respondent in dollars",
  cex.lab = 0.7)
```



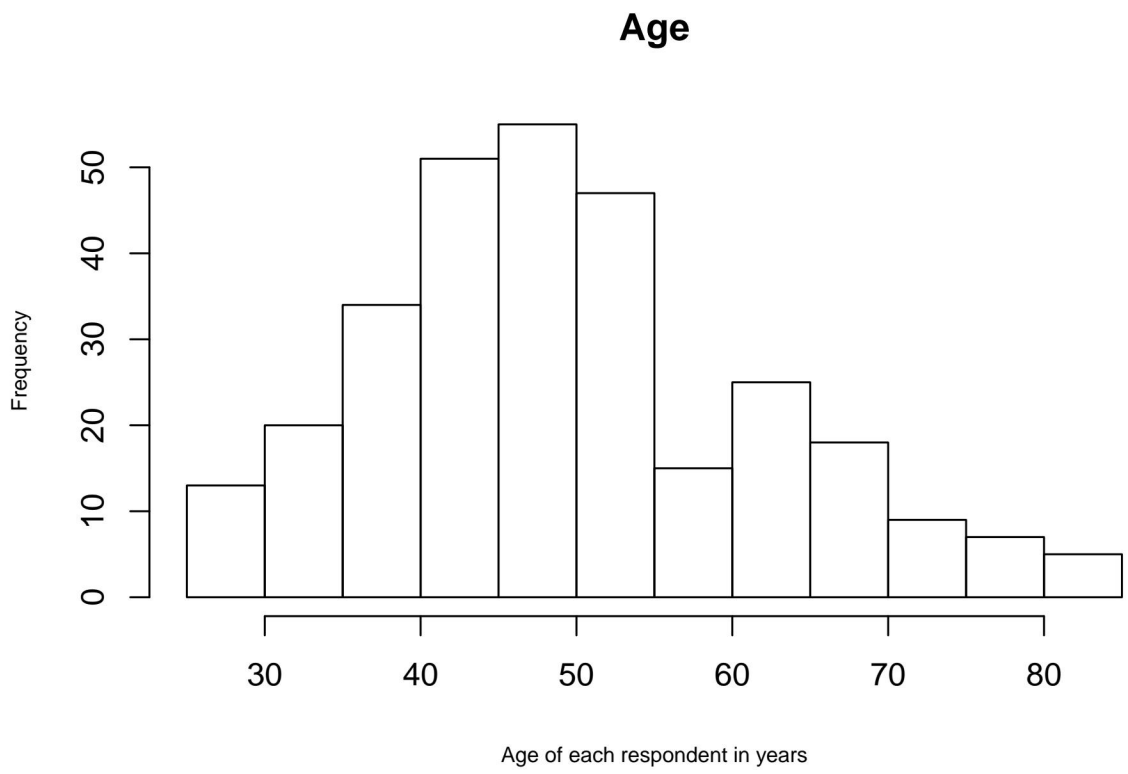

```
summary(nyc$Income)
```

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

Summary: The distribution of `Income` is unimodal and skewed right; this is supported by the quantitative summary, as the mean (42266 dollars) is greater than the median (39000 dollars). The minimum income is 1440 dollars and the maximum number of images is 98000 dollars. The distribution does appear to have outliers.

EDA on X (Age, NYCMove, MaintenanceDef)

```
hist(nyc$Age,
  main = "Age",
  xlab = "Age of each respondent in years",
  cex.lab = 0.7)
```

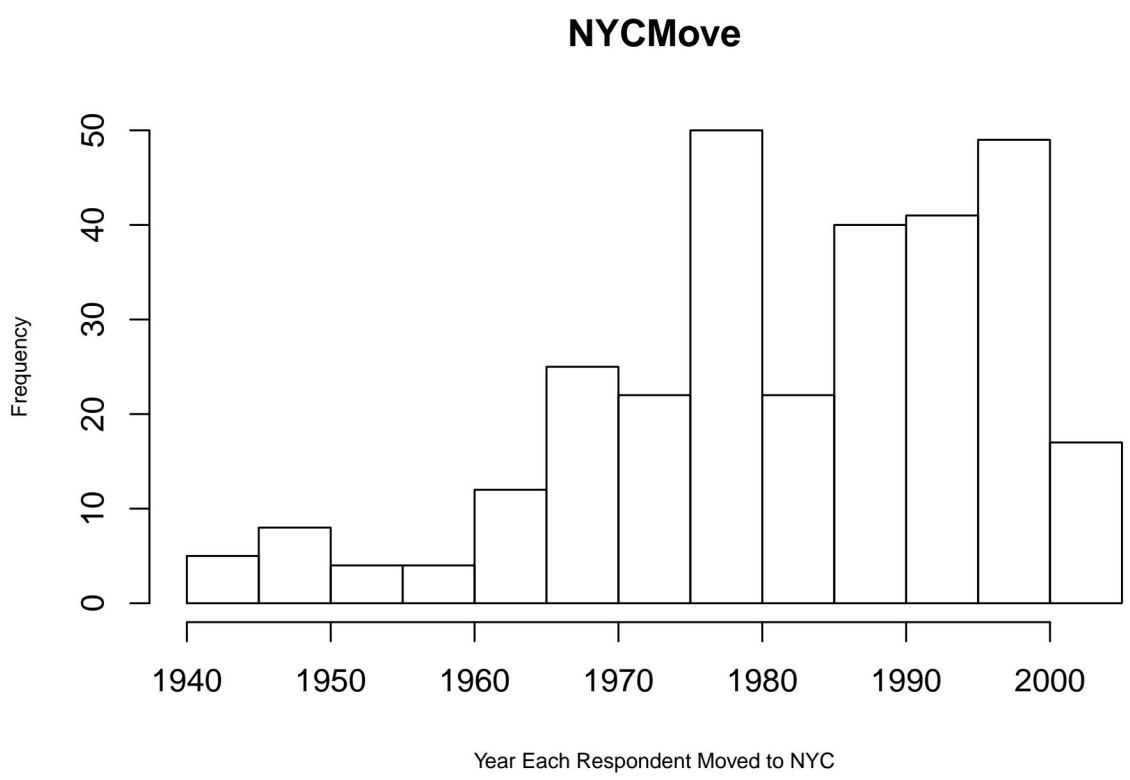



```
summary(nyc$Age)
```

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

Summary: The distribution of `Age` is unimodal and skewed right; this is supported by the quantitative summary, as the mean (50.03 years) is greater than the median (49 years). The minimum number age is 26 and the maximum age is 85. The distribution does appear to have outliers.

```
hist(nyc$NYCMove,
  main = "NYCMove",
  xlab = "Year Each Respondent Moved to NYC",
  cex.lab = 0.7)
```

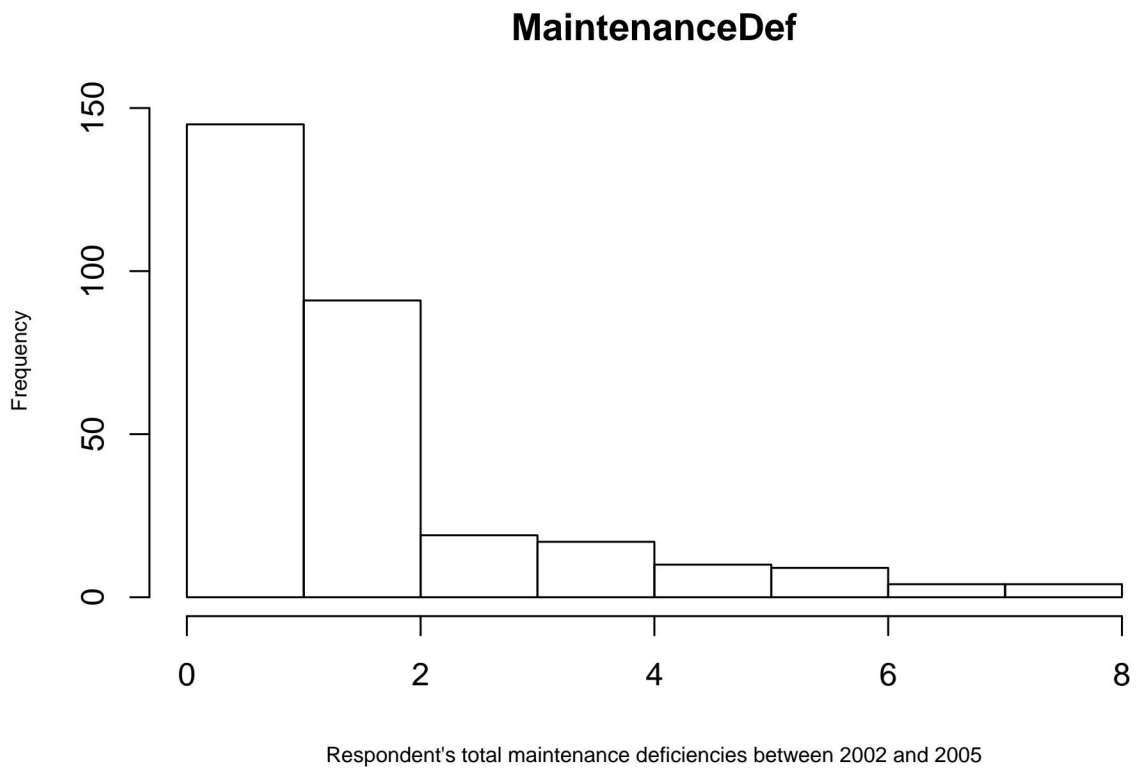



```
summary(nyc$NYCMove)
```

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

Summary: The distribution of `NYCMove` is unimodal and skewed left; this is supported by the quantitative summary, as the mean (1983) is less than the median (1985). The earliest year moved is 1942 and the latest year moved is 2001. The distribution does appear to have outliers.

```
hist(nyc$MaintenanceDef,
  main = "MaintenanceDef",
  xlab = "Respondent's total maintenance deficiencies between 2002 and 2005",
  cex.lab = 0.7)
```




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

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

Summary: The distribution of `MaintenanceDef` is unimodal and skewed right; although the quantitative summary shows the mean (1.98 deficiences) is less than the median (2 deficiences), this is explained by the large amount of '0' values for maintenance deficiencies. The minimum number of deficiences is 0 and the maximum number of deficiencies is 8. The distribution does appear to have outliers.

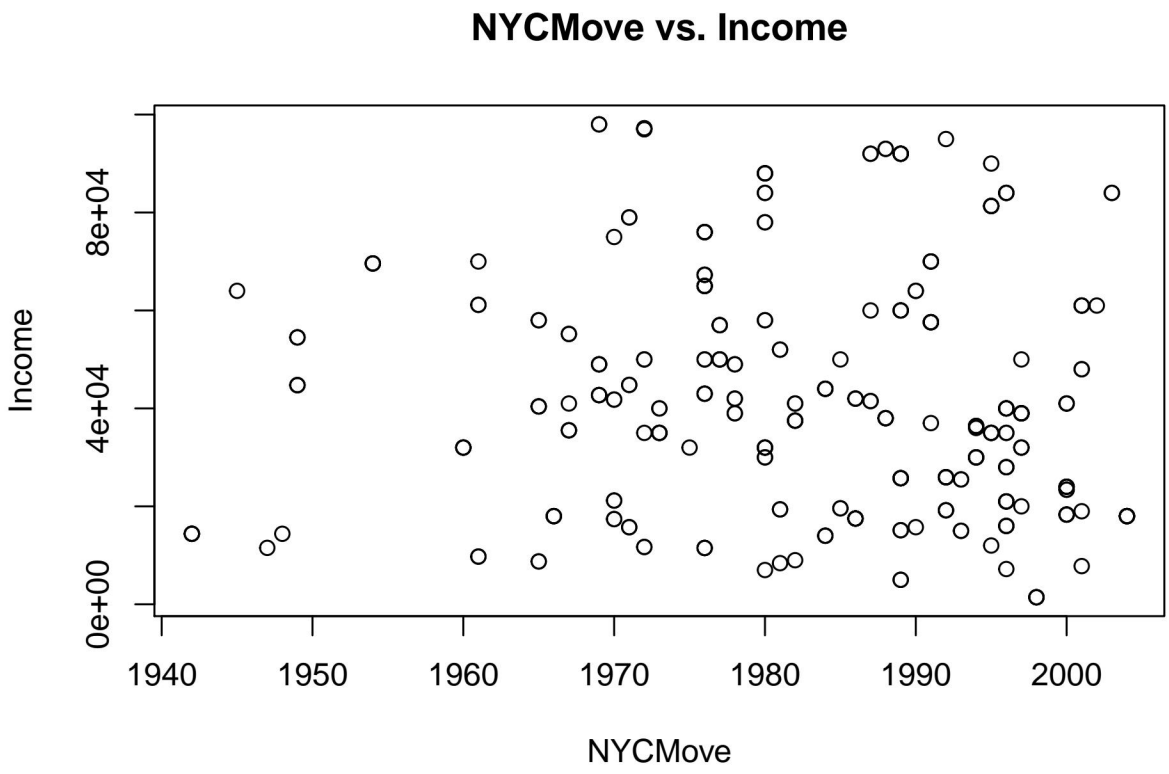
Bivariate EDA

Because we have finished the univariate data analysis, we can now look at the relationship between the response variable and each predictor variable using scatterplots.

```
plot(Income ~ Age,
  data = nyc,
  main = "Age vs. Income",
  xlab = "Age",
  ylab = "Income")
```

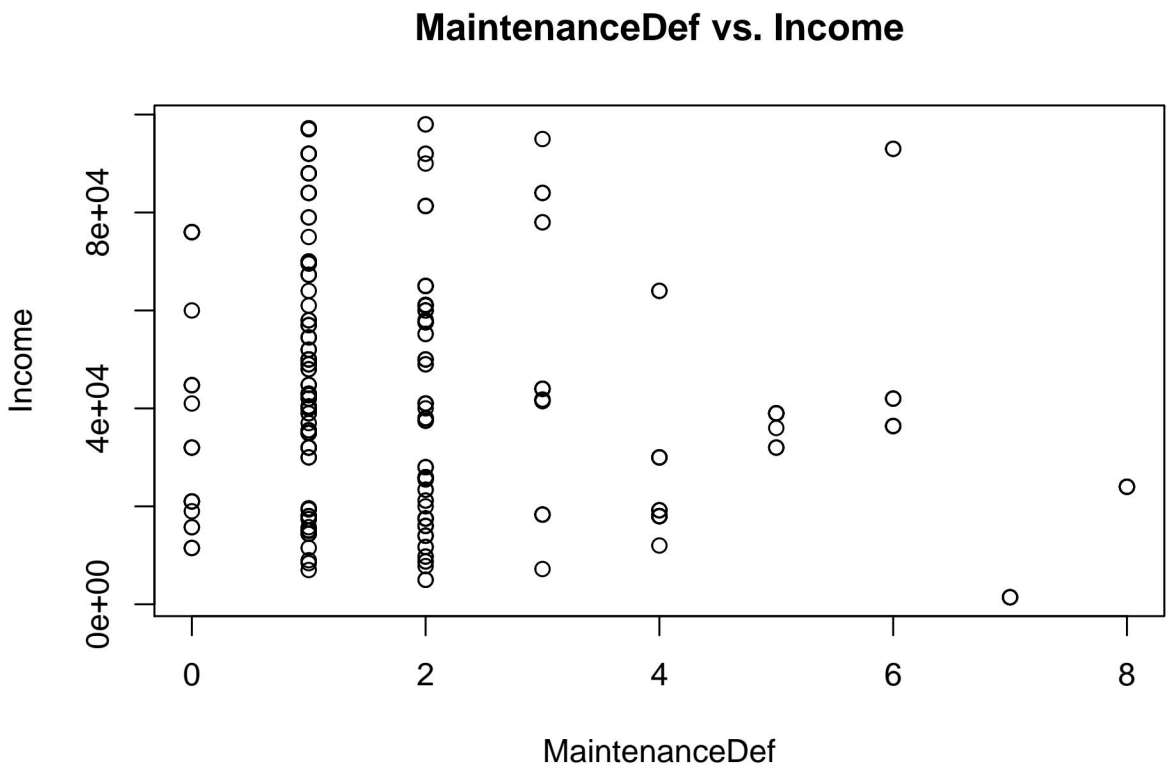
Summary: The values seem to be pretty spread out throughout the scatterplot.

```
plot(Income ~ NYCMove,
  data = nyc,
  main = "NYCMove vs. Income",
  xlab = "NYCMove",
  ylab = "Income")
```



Summary: The values seem to cluster around the right side of the scatterplot.

```
plot(Income ~ MaintenanceDef,
  data = nyc,
  main = "MaintenanceDef vs. Income",
  xlab = "MaintenanceDef",
  ylab = "Income")
```



Summary: The reason why the points are in columns is because the maintenance deficiences are discrete. The values seem to cluster around the left part of the scatterplot.

Modeling

Now that we have completed the univariate data analysis and bivariate data analysis, we can begin creating the regression model.

Below, we created a regression model using all of the variables (none of which were transformed).

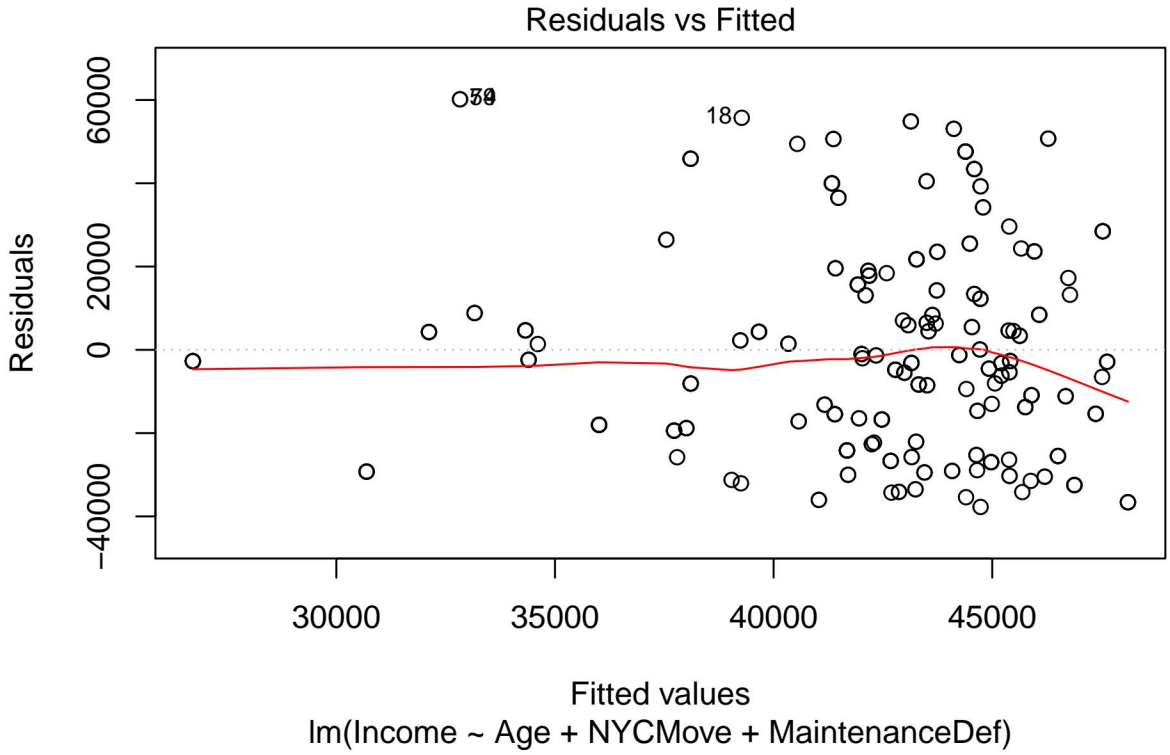
```
nyc.original.mod <- lm(Income ~
  Age + NYCMove + MaintenanceDef,
  data = nyc)

summary(nyc.original.mod)
```

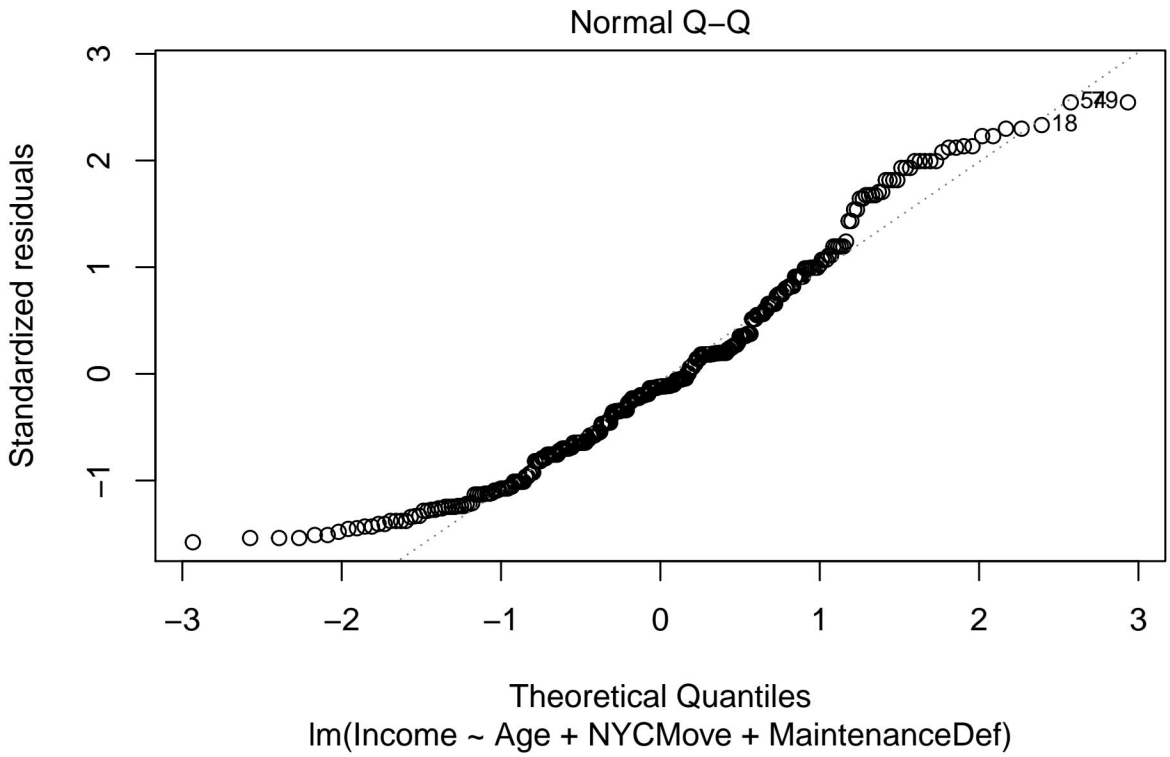
```
##
## Call:
## lm(formula = Income ~ Age + NYCMove + MaintenanceDef, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37734 -18010  -2878  14971  60171
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    237408.41  278939.01   0.851   0.3954
## Age               -71.98     144.97  -0.496   0.6199
## NYCMove           -94.34     138.82  -0.680   0.4973
## MaintenanceDef  -2273.22     964.72  -2.356   0.0191 *
## ---
## 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
```

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




```
plot(nyc.original.mod, which = 2)
```



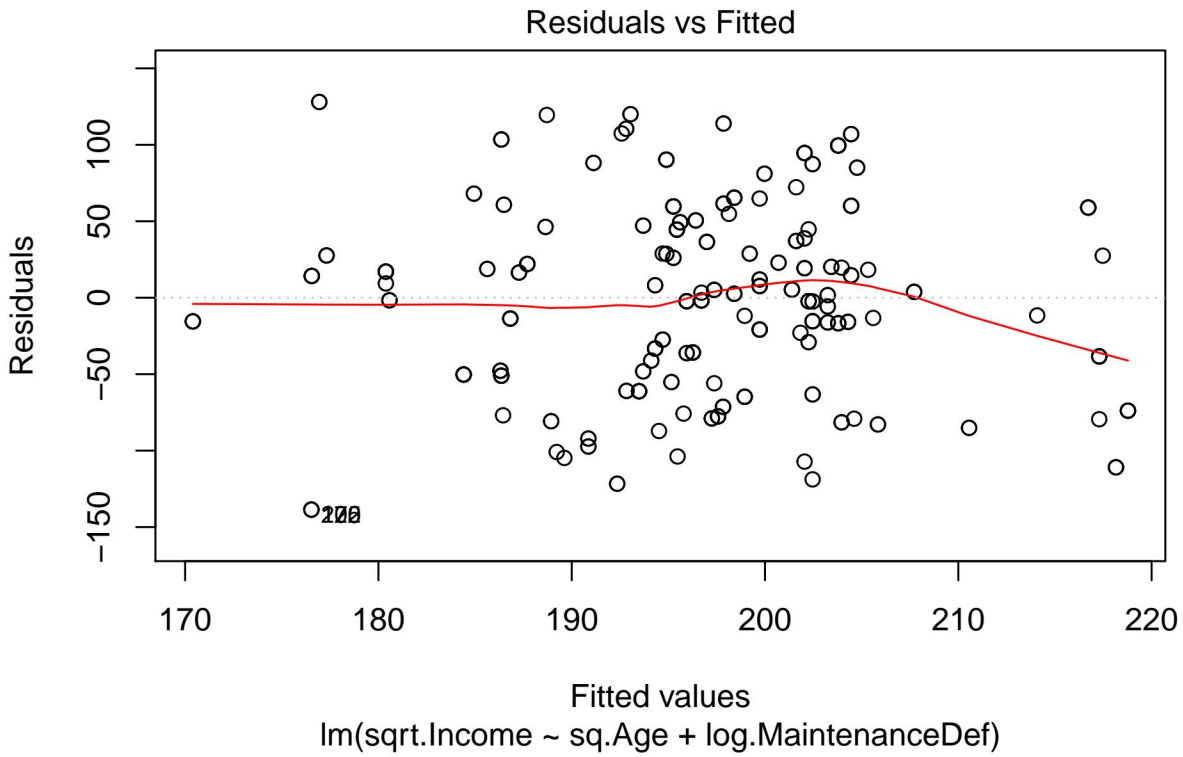
From the above summaries and diagrams, we see that this model has problems. While the `MaintenanceDef` variable is significant, the diagnostics for this model could be better. Specifically, the residual plot has more "sparseness" on the left side, and the residuals seem to be clustered around the right side. Additionally, the ends of the qqplot we see deviation, especially on the positive theoretical quantiles. Thus, we made the decision to transform some of the variables and try new models to find better diagnostics.

After trying out different ways to improve the diagnostics, we decided for our final model to take the square root of `Income`, square `Age`, and take the log() of `MaintenanceDef`. The diagnostics and summary are below.

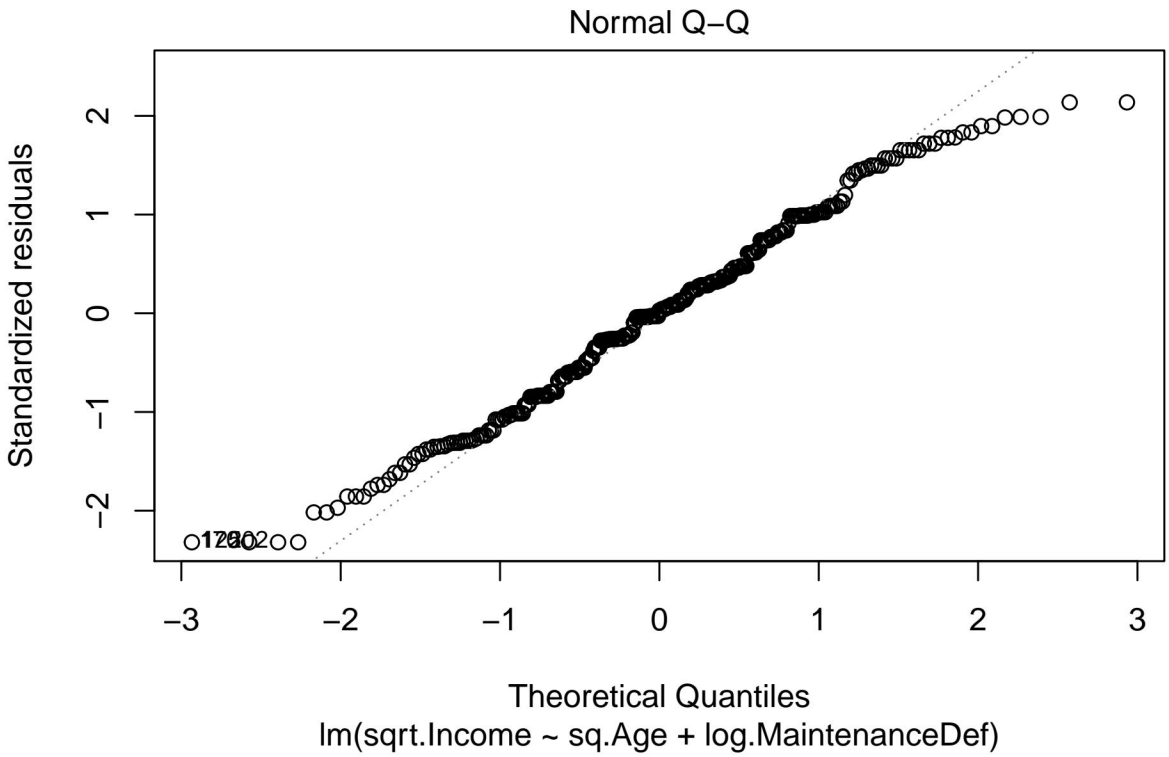
```
nyc$sqrt.Income <- (nyc$Income)^0.5
nyc$sq.Age <- (nyc$Age)^2
nyc$shift.MaintenanceDef <- (nyc$MaintenanceDef) + 1.1
nyc$log.MaintenanceDef <- log(nyc$shift.MaintenanceDef)
```

```
nyc.mod <- lm(sqrt.Income ~
  sq.Age + log.MaintenanceDef,
  data = nyc)
```

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




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



By looking at the qqplot and residual plot, we see that these diagnostics with the transformed variables have improved in comparison to the original model.The residual plot is random and does not seem clustered to one particular side. The qqplot still deviates on the ends but much less so than the qqplot on the original model.

As a result, the results of the residual plot (of which the residuals are shown to be random and distributed relatively equally above the zero 0) lead us to assume constant spread, mean zero, and independence; this was the best residual plot created out of all other attempted models. The results of the qqplot demonstrate that the normality condition is not invalidated, as the deviations on the end are not enough to justify invalidating the normality condition because otherwise the points are close to or on the line; this was the best qqplot created out of all other attempted models.

```
car::vif(nyc.mod)
```

```
##             sq.Age log.MaintenanceDef
##           1.073502           1.073502
```

As seen, none of the vifs are above 2.5, so the we can assume that that our predictor variables are weakly or not correlated. In other words, our multicollinearity assumption has been met.

Below, we have the final model summary.

```
summary(nyc.mod)
```

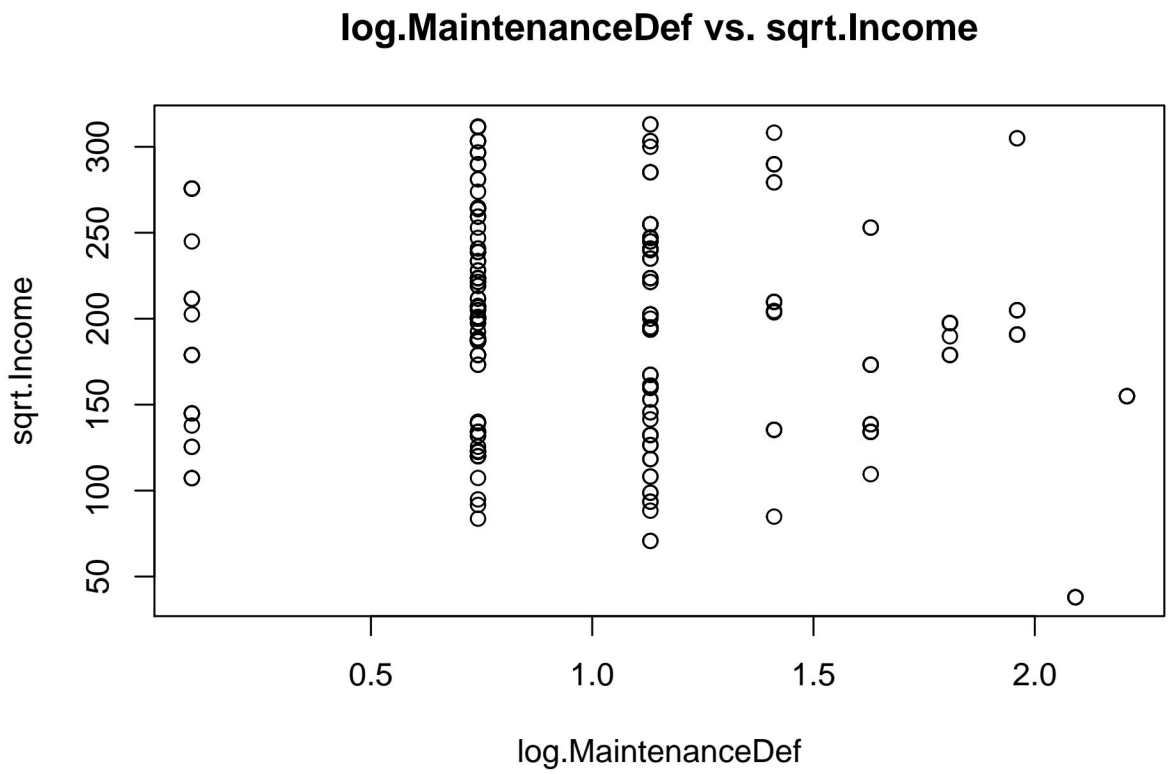
```
##
## Call:
## lm(formula = sqrt.Income ~ sq.Age + log.MaintenanceDef, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -138.584  -47.958    1.695   44.561  128.017
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         223.459558  12.466135  17.925  < 2e-16 ***
## sq.Age               -0.002038   0.002701  -0.754  0.45123
## log.MaintenanceDef -21.436033   7.789544  -2.752  0.00629 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.43 on 296 degrees of freedom
## Multiple R-squared:  0.02495,    Adjusted R-squared:  0.01836
## F-statistic: 3.787 on 2 and 296 DF,  p-value: 0.02377
```

We consider this a reasonable model for predicting `Income`. As seen, predictor coefficient for `log.MaintenanceDef` is significant. While the predictor coefficient for `sq.Age` is not significant, the predictor variable was kept in the model because it noticeably improved the quality of the residual plot and qqplot. This was also the most reasonable model produced because it had the highest R^2 in comparison to other attempted model. The model itself is significant because the p-value from the F-test is 0.02377, which is less than 0.05.

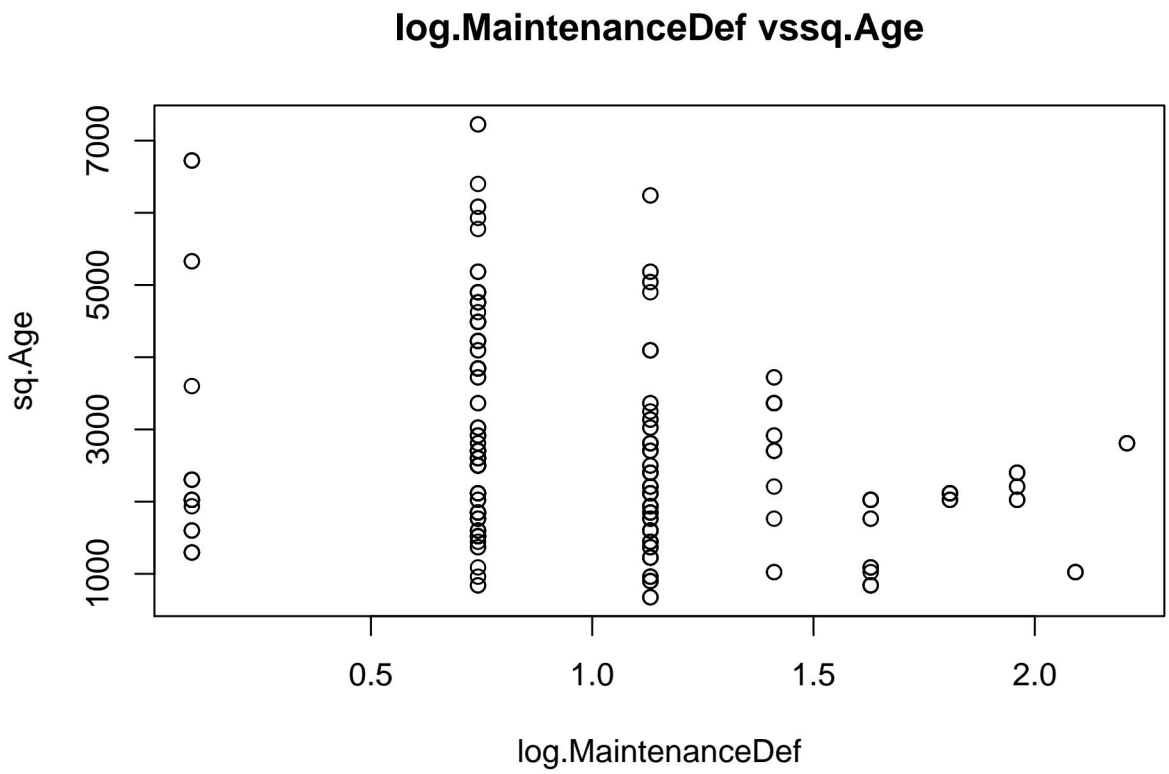
The other models attempted included combinations of all variables, both transformed and not transformed. Specifically, we tried different combinations of variables with no transformation, squared, square root, and log(). Out of these models created, this model produced the the most reasonable balance between R^2, significant predictor coefficient, and residual plot/qqplot.

```
plot(sqrt.Income ~ log.MaintenanceDef,
  data = nyc,
  main = "log.MaintenanceDef vs. sqrt.Income",
  xlab = "log.MaintenanceDef",
  ylab = "sqrt.Income")
```



From the above plot showing `sqrt.Income` vs `log.MaintenanceDef`, we see that there is a slight positive trend, except for an outlier on the bottom right of the scatterplot. Thus, we demonstrate a relatively linear relationship between these two quantitative variables and therefore justifying the linearity condition.

```
plot(sq.Age ~ log.MaintenanceDef,
  data = nyc,
  main = "log.MaintenanceDef vssq.Age",
  xlab = "log.MaintenanceDef",
  ylab = "sq.Age")
```



Based on this scatterplot, there does not seem to be any interaction between the predictor variables. We do not use an interaction plot because all of our predictor variables are quantitative.

Prediction

We now have a justified model, so we can begin our prediction. Specifically, we are interested in predicting the income for a household with three maintenance deficiencies and whose respondent's age is 53 and who moved to NYC in 1987.

```
Age <- 53
MaintenanceDef <- 3
sqrtIncome <- -0.002038*Age^2 + -21.436033*log(MaintenanceDef) + 223.459558
Income <- sqrtIncome^2
Income
```

```
## [1] 37707.79
```

The predicted income of a 53 year old respondent with three maintenance deficiencies in 1987 is $37707.79. This is less than both the median and mean income from the sample.

Discussion

When we conducted this analysis, we chose to predict `Income` using the predictor variables `Age` and `MaintenanceDef`. We did not include the predictor variable `NYCMove` because the models using that variable

were not as strong (in terms of residual diagnostics, significant predictor coefficient) as the final model. However, we should make note of one limitated: `Age` is not a significant predictor coefficient, although it did help our model diagnostics. In other words, the model created only had one significant predictor coefficent, which was `MaintenanceDef`; according to our model, `MaintenanceDef`, or the number of maintenance deficiencies, is the only significant predictor for `Income`. Thus, we only had one significant coefficient predictor in our final model.

Now, we will compare our final model to the initial beliefs in the scenario. We did not find that older respondents have higher income in our EDA, and we did find that maintenacne deficiences correspond to income (as shown in our final model). We did, however, show in our final model that we did not consider the relationship between the year of moving to NYC and the household income beneficial to our model and never found a significant coefficient predictor for it during our other attempts.

We noted one outlier in the `sqrt.Income` vs `log.MaintenanceDef` scatterplot, so it may be helpful to seek out that individual for further information.

We note that even within sub-boroughs, there are differences within housing facilities and demographics of residents. We would like to learn more about the different areas within these sub-boroughs for better understanding of the relationship between income and housing conditions.

It is in our best interest to continue analyzing the data from the survey in years to come. It would also be interesting if we compared the survey data now to that in years past. As a result, we would be able to see how "the other half" and their living conditions and demographics have changed over time.

In all, through this paper, we can get a better understanding of the demographics and the living conditions of "the other half" from the survey data, as well as the relationship between `Income` and other predictor variables in our final model.

1 whole project / all pages (please tag all the pages in gradescope when you upload) **98 / 98**

✓ **- 0 pts** **Correct**

INTRO: 10/10 pts
Concise and well-written

EDA: 25 pts, as follows:
---
* EDA on Y: 3/3 pts
---
* EDA on each X: 9/9 pts total
---
* EDA on relationship between Y and each X: 9/9 pts total
---
* EDA other: 4/4 pts
Overall good style and formatting

MODELING: 38 pts, as follows:
Good; all details included and appropriate model selection
---
* residual plot: 4/4 pts
---
* discuss residual plot: 6/6 pts
---
* qqplot: 4/4 pts
---
* discuss qqplot: 3/3 pts
---
* discuss linearity assumption between Y and Xs: 4/4 pts
---
* discuss interactions: 3/3 pts
---
* discuss multicollinearity: 3/3 pts
---
* other valid reasons for choice of model: 2/2 pts
---
* showing final model summary: 5/5 pts

---

* discussing significance: 4/4 pts

PREDICTION: 15/15 pts

CONCLUSION: 10/10 pts
