Understanding the relationship between Household Income and various demographic and housing measurements

Jae Kang
jkang2
36-401

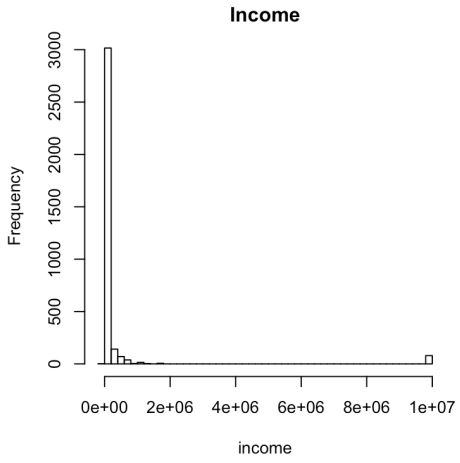
Introduction :

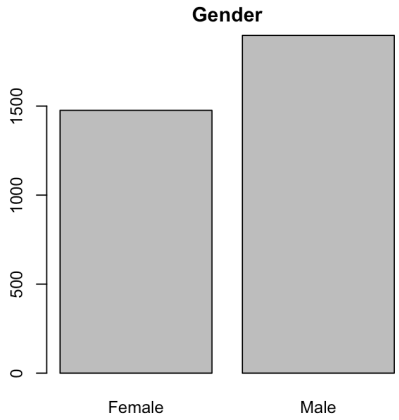
This report will examine the relationship between household income and various demographic and housing quality measurements taken from the New York Housing and Vacancy Survey done every three years. **In particular, this report will cover two hypotheses about the difference in Caucasian households and Hispanic households considering that all the other housing and demographic variables are constant and also the effect that water leakage may have on the relationship between age and household income (Question 1).** It is crucial to understand the correlation between various housing quality factors and household income because the New York City government primarily needs to accurately predict rent vacancy rates for changes in rent regulation laws.

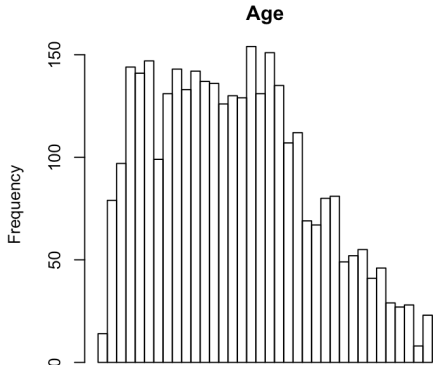
Exploratory Data Analysis :

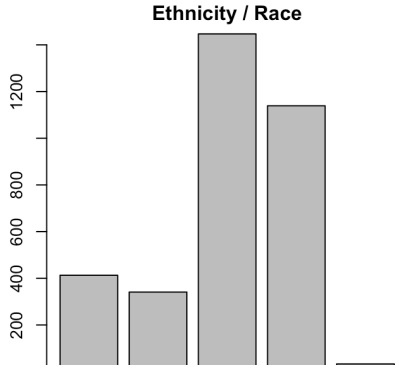
Firstly, there were **blaring outliers in the data where the household incomes were either in the millions or in the negatives (Question 2)**. This could also be observed in the multivariate EDA section where we plotted the box plots between variables such as the ethnicity categorical variable and household income. Thus it was diagnosed that these outlier data points would not contribute to the model fit and will be removed in the initial modeling phase.

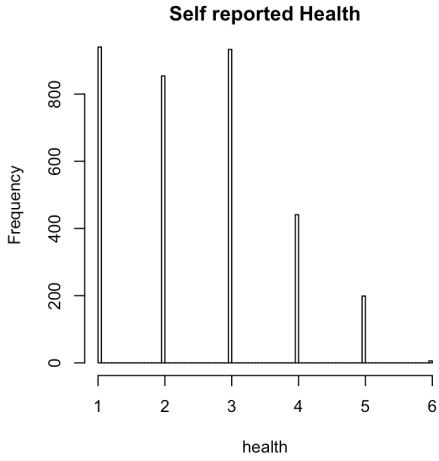
However, in the initial data there are **3373 households that have been sampled as instances for this survey** and with regards to ethnicity the **overwhelming majority of them are either Caucasian or Hispanic. The age distribution is pretty evenly distributed from the mid twenties onwards until the mid fifties** where the number of households with ages above 60 start to decline **(Question 2)**.

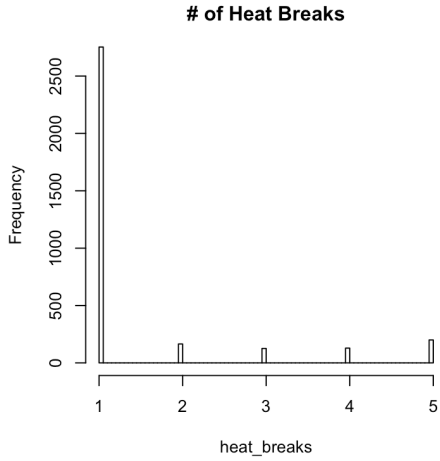

Income
Frequency
0
500
1000
1500
2000
2500
3000
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
income

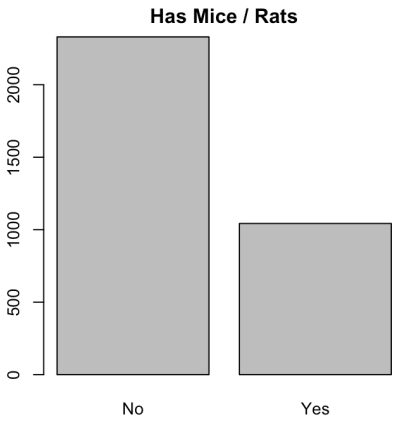

Gender
0
500
1000
1500
Female
Male

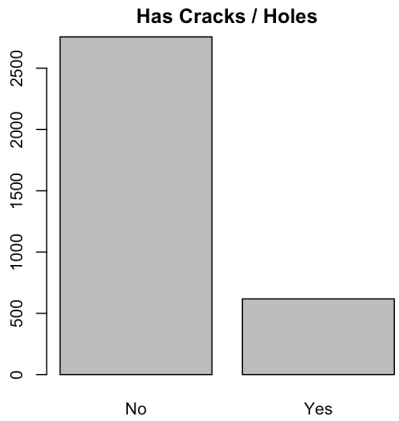

Age
Frequency
0
50
100
150

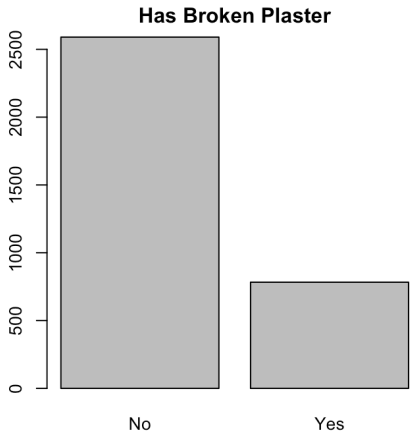

Ethnicity / Race
200
400
600
800
1200

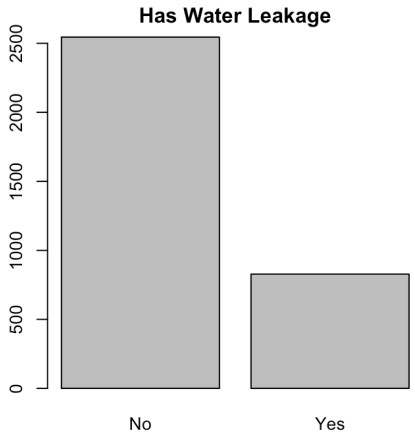

Self reported Health
Frequency
0
200
400
600
800
1
2
3
4
5
6
health


# of Heat Breaks
Frequency
0
500
1000
1500
2000
2500
1
2
3
4
5
heat_breaks

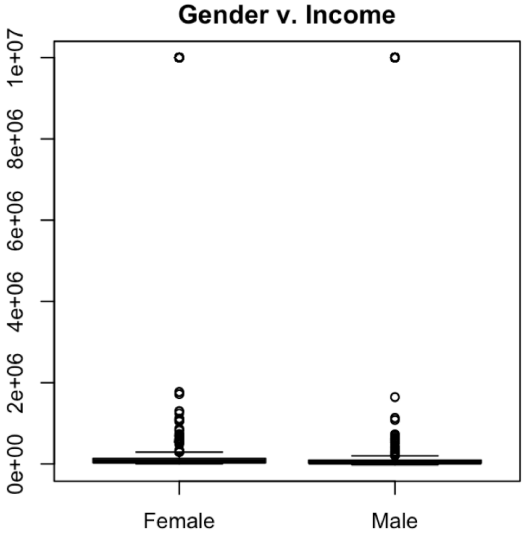

Has Mice / Rats
0
500
1000
1500
2000
No
Yes

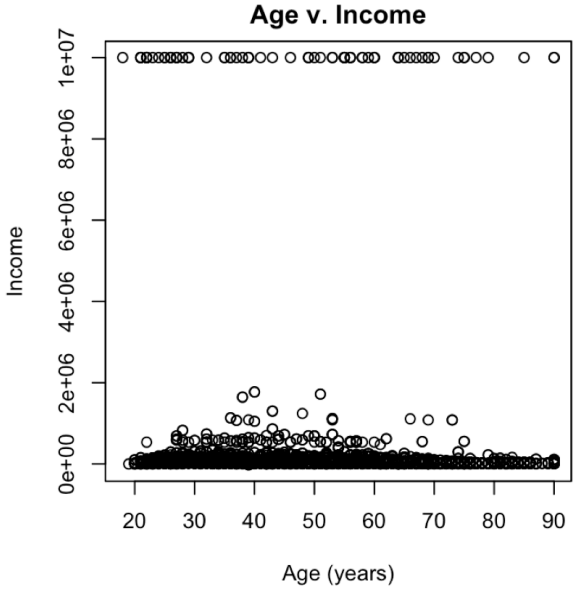

Has Cracks / Holes
0
500
1000
1500
2000
2500
No
Yes

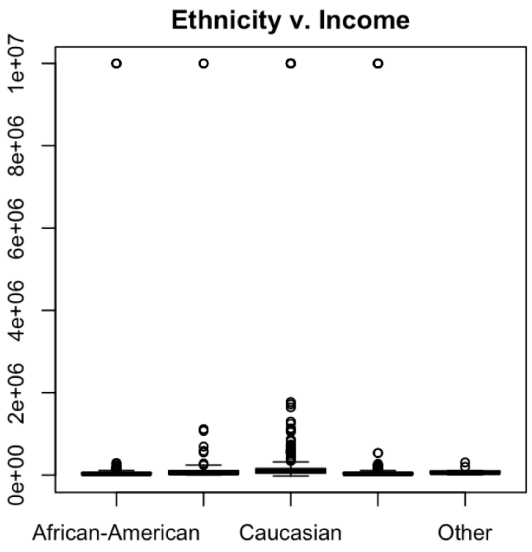


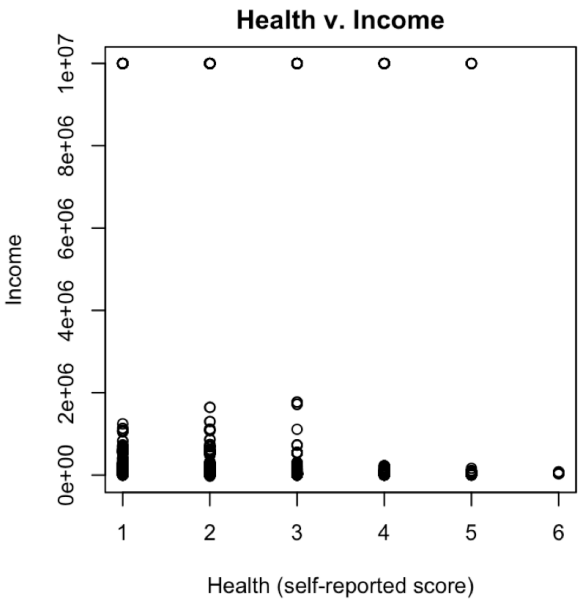


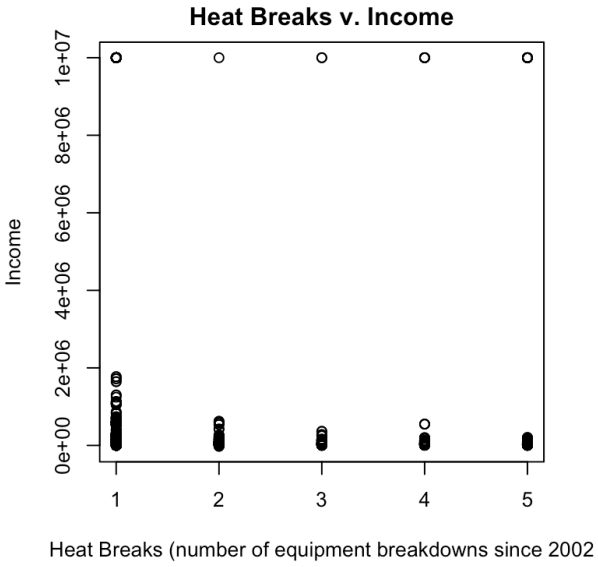
The multivariate EDA process also showed the blaring outliers that skewed the box plots and scatter plots in an extreme manner (Question 3). Not much could be observed from the multivariate EDA process because the skewed nature of these box plots and scatterplots but these outliers will be dealt with (through data cleaning) in the next phase of the report.

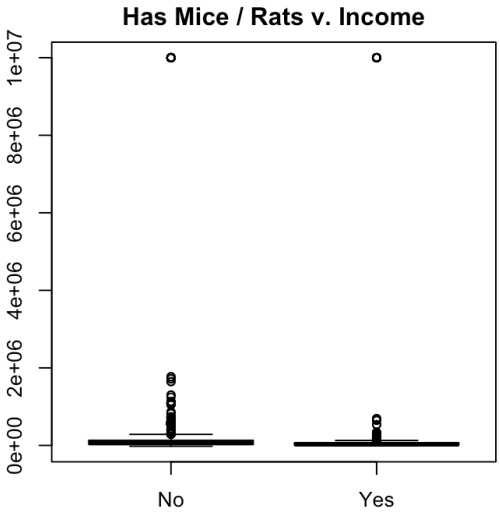


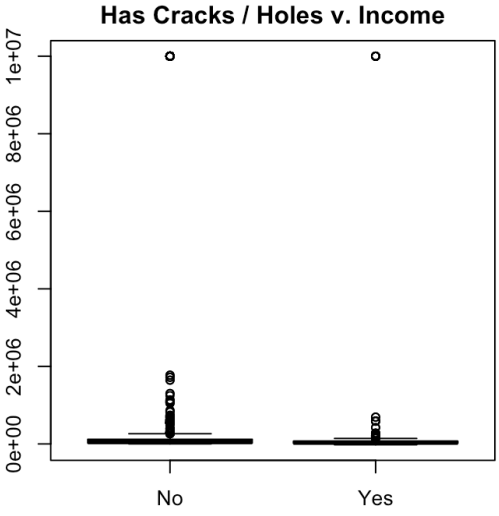


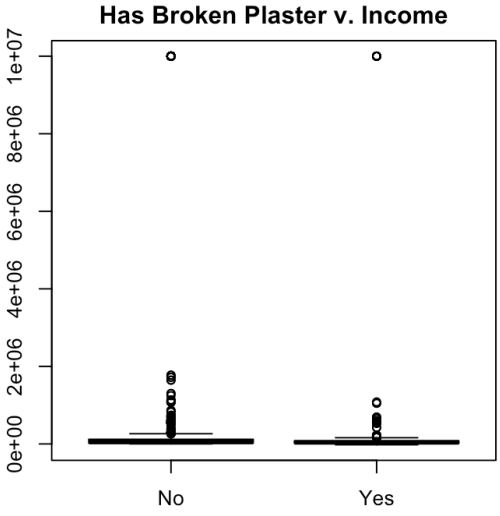

Ethnicity v. Income
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
African-American
Caucasian
Other


Health v. Income
Income
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
1
2
3
4
5
6
Health (self-reported score)


Heat Breaks v. Income
Income
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
1
2
3
4
5
Heat Breaks (number of equipment breakdowns since 2002


Has Mice / Rats v. Income
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
No
Yes

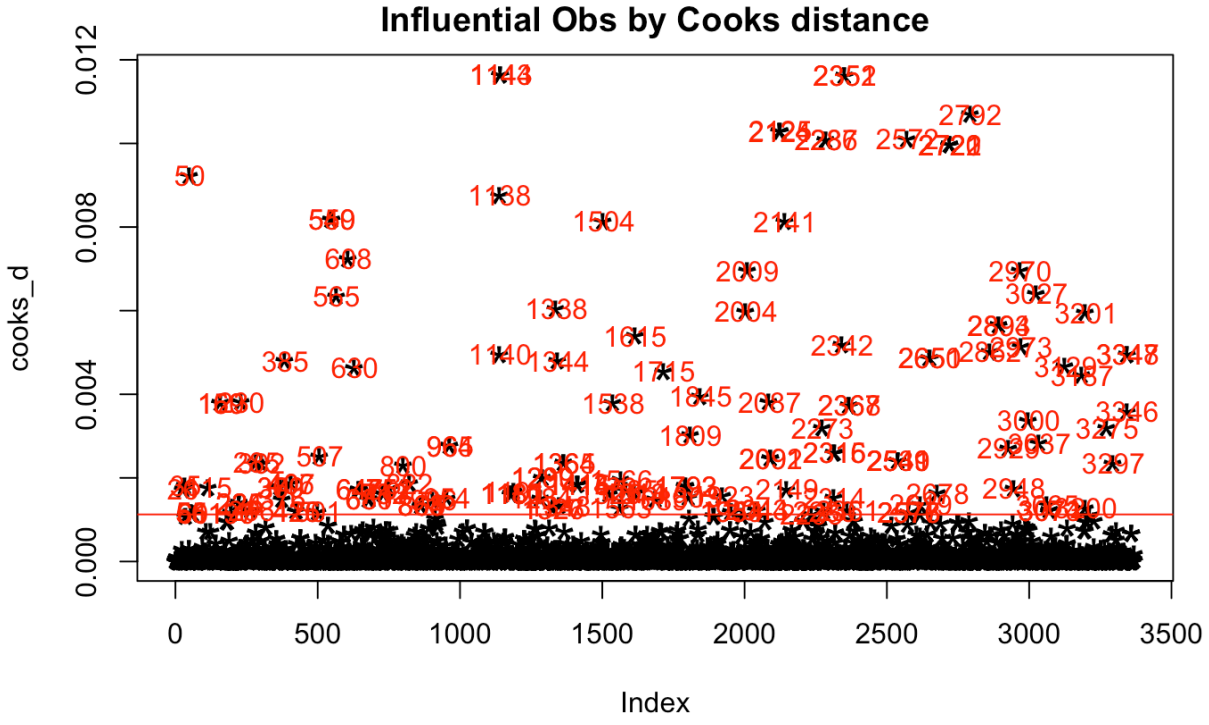

Has Cracks / Holes v. Income
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
No
Yes


Has Broken Plaster v. Income
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
No
Yes

Initial Modeling & Diagnostics :

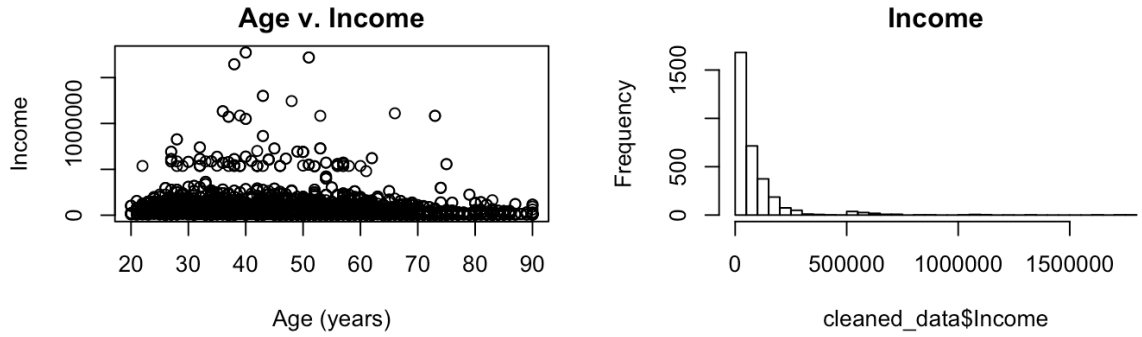
In the initial modeling phase, **we first fit the multiple linear regression model with a log transformation on income with all the covariates considered. The covariates treated as discrete categorical variables were Ethnicity (because there were several categories of Ethnicity that could obviously not be considered as numeric values), WaterLeakage, BrokenPlaster, CracksHoles, Gender, and MiceRats (Question 4)**. We will test the very basic multiple regression model first and add or remove terms as necessary. **We will log transform income because income tended to be right skewed from our data and we want to observe the values more frequent in the lower range (Question 7)**. (Model 1)

Note that there are truncated values and NaN values produced because of the **log transformation on the income response variables that might have produced negative log values (negative incomes in our dataset). These values will be dropped along with the data points deemed as outliers or not contributing to the model (Question 2, 8)**. We will deem data points outliers by using Cook's distance as a standard and we will remove all the data points that are 4 times the mean Cook distance of all the data.



154 data points deemed as either outliers (by the average Cook distance measurement) or truncated data were removed because they did not add

anything specific to the multiple regression model we were attempting to fit (Question 2, 8). These values have a tendency to skew the regression line towards these outliers or truncated data points that don't tell us anything meaningful about the pattern or relationship between the housing covariates and income.



After cleaning the data and removing the outliers/missing values, the histogram for income looks more evenly distributed and the scatterplot for Age v. Income is not as skewed. Therefore, we can conclude that the data has been cleaned effectively and removed of data that may affect our model negatively.

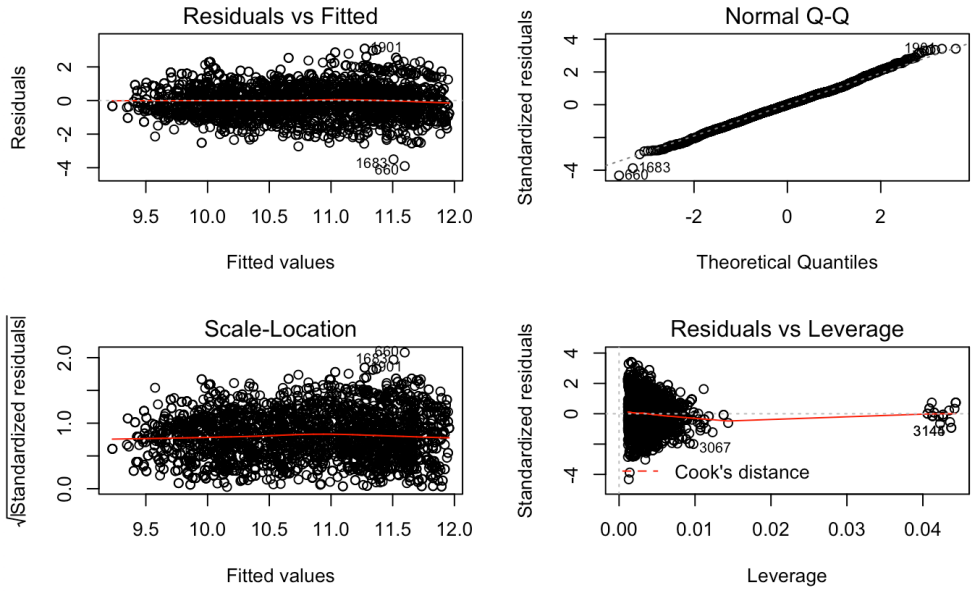
In addition, for the categorical Ethnicity variable **the reference point was changed to Caucasian so that it would be easier to test** the hypothesis that the average household income is different for Caucasian and Hispanic households **(Question 4)**.

Now in our new multiple linear regression model fitted with the cleaned data (Model 2), we can observe in the summary that there are several coefficient estimates with non-significant p-values at the 0.05 level. These would be CracksHoles, BrokenPlaster, MaintenanceDef which have p-values of 0.4074, 0.3905 and 0.8295 respectively (Figure 1). Therefore, we want to **test the null hypothesis that all these slope coefficients for the variables mentioned above are zero. We will do this through a partial F-test using the anova table and a null model fit without the variables CracksHoles, BrokenPlaster and MaintenanceDef (Question 6)**.

Note that in our null model without the variables that we are testing as zero, the p-value for the F statistic is shown as much greater than 0.05 (Figure 2), thus at the 0.05 significance level, we cannot reject the null hypothesis that the slope coefficients for CracksHoles, BrokenPlaster and MaintenanceDef are zero. Therefore, we will interpret this information as the fact that fitting the slopes for the CracksHoles, BrokenPlaster and MaintenanceDef variables as non-zero will not reduce the mean-squared error of the model more than we would expect by just random noise and thus this tells us that they do not contribute much to the prediction of household income.

Therefore, we will safely use this assumption to delete CracksHoles, BrokenPlaster and MaintenanceDef from our model (Question 6).

In addition, **to test the second hypothesis that the relationship between age and household income is different depending on whether or not water leakage has occurred in the apartment, (for households whose other variables are constant), we must include an interaction variable and slope coefficient fit to test this efficiently (Question 5).** Thus, we will also fit a model with the interaction variable between age and water leakage
(Model 3).



Testing the assumptions of our model (Model 3), we can observe that by our residuals plot that the **linearity assumptions hold because our residuals v. fitted values plot is centered around zero**. In addition, **homoscedasticity holds because there are no extreme outliers and the scattered plots are fairly balanced.** Also since there is **no pattern across observations from the abline in the residuals v. fitted values and we can assume that the observations are independent and identically distributed** across the data. Finally, notice that due to the qqplot's points being consistent with the line, **the Gaussian noise assumptions hold for this model as well (Question 9).**

Results :

Using Model 4,

The First Test (Question 10) :

Null Hypothesis :

The expected difference between average Caucasian household incomes and average Hispanic household incomes are the same when all the other demographic measurements and housing quality characteristics are constant.

Alternate Hypothesis :

The expected difference between average Caucasian household incomes and average Hispanic household incomes are different when all the other demographic measurements and housing quality characteristics are constant.

Test statistic : the fitted slope coefficient for the Ethnic variable in cleaned_model

Note that Caucasian is the reference point so we just need to observe the Hispanic row, and when we do observe cleaned_model's EthnicHispanic's p-value, it is significant at the 0.05 level. Therefore, we can reject the null hypothesis and confidently state that the expected difference between average Caucasian household incomes and average Hispanic household incomes are different.

The Second Test (Question 10):

Null Hypothesis :

The expected difference that an extra year in age will have on household income is the same depending on whether or not a water leakage has occurred in the apartment, assuming that all other household characteristics are the same.

Alternative Hypothesis :

The expected difference that an extra year in age will have on household income is different depending on whether or not a water leakage has occurred in the apartment, assuming that all other household characteristics are the same.

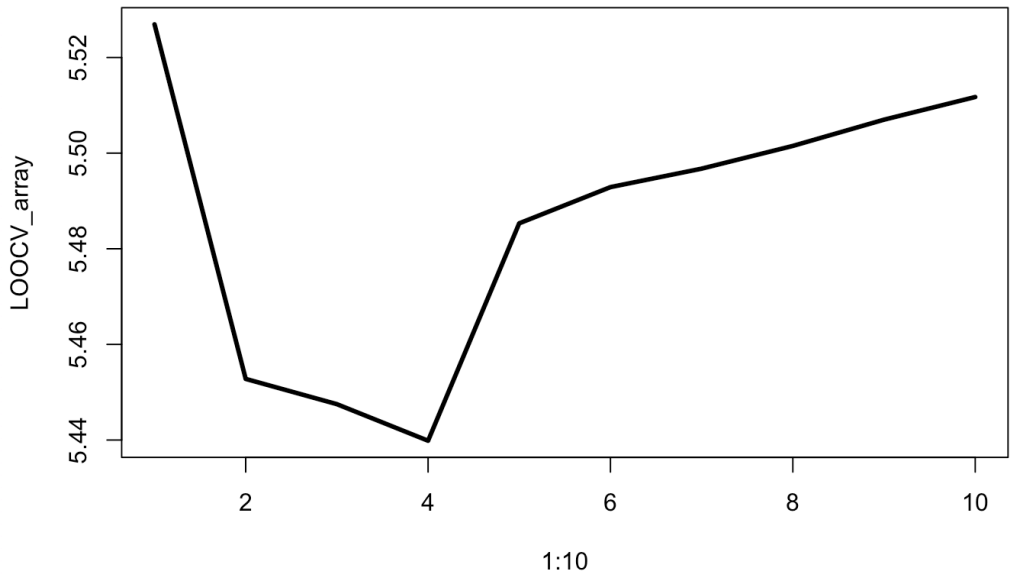
Test statistic : the fitted slope coefficient for the interaction variable between Age and WaterLeakage.

However, observe that the interaction variable for age and water leakage has a insignificant p-value at the 0.05 level (0.99395) and thus we cannot reject the null hypothesis that the effect that a water leakage may have on the relationship between age and household income is zero. Therefore, we will assume that the water leakage and age interaction variable's slope coefficient is negligible and will omit it in the model.

In addition, we can observe that WaterLeakageYes has a insignificant p-value of **0.47977 which is greater than then 0.05 significance level and thus should have a zero association with household income (Question 6)** if we set the null hypothesis such that slope coefficient of the WaterLeakage variable is zero. Thus, we will add the model with WaterLeakage omitted in our pool of models to consider (Model 4).

Now, for our model selection phase, our pool of models will include the very first multiple linear regression model fit (Model 1), the outlier removed model including CracksHoles, Maintenance_Def, BrokenPlaster (Model 2), the outlier removed model with WaterLeakage and the interaction effect between Age and WaterLeakage and the 3 unnecessary indicator variables mentioned above removed (Model 3), the final fit multiple linear regression model without WaterLeakage (Model 4), and the polynomial models with Age raised up to the power of 7 (Models 5 to 10).

Then we plot the models (as indexed in order from 1 to 10) against their LOOCV score below (Question 11).



Note that Model 4 has the lowest LOOCV score and thus has the lowest generalization error. **Therefore, the multiple linear regression model with age, gender, ethnicity, health, heatBreaks, and miceRats predicting log transformed household income**

fit to the cleaned data will be chosen as the best model of all these predictive models that also meets all necessary model assumptions (Question 7, 11).

Conclusion :

It was concluded that there exists a nonzero association between a change in age with the expected difference in income given all other variables constant. **One possible reason for this is that Caucasian households may tend to have better resources or education and thus may be a better predictor of the household income (Question 12).**

It was also concluded that the relationship between age and income differing depending whether a water leakage occurred or not given all other variables are constant was deemed to not hold and thus we removed the interaction effect between Age and WaterLeakage in our final model (Model 4). **One possible reason for this is that whether or not a water leakage occurs in a home, it doesn't change anything about the age and income of the household (Question 12).**

Our final model chosen was Model 4 with no polynomial variables and no interaction terms included under our LOOCV score criterion. **This may be due to quadratic or other polynomial models overfitting the data and thus not generalizing well to new instances and also with unnecessary housing or demographic factors not adding significant predicting power to predicting the household income (Question 12).**

Appendix :

Figure 1:

```
Call:
lm(formula = log(cleaned_data$Income) ~ cleaned_data$Gender +
    cleaned_data$Age + cleaned_data$Ethnic + cleaned_data$Health +
    cleaned_data$HeatBreaks + cleaned_data$MiceRats + cleaned_data$CracksHoles +
    cleaned_data$BrokenPlaster + cleaned_data$WaterLeakage +
    cleaned_data$MaintenanceDef)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9026 -0.5770  0.0005  0.5968  3.0836

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      12.391413   0.059923 206.787  < 2e-16 ***
cleaned_data$GenderMale          -0.259582   0.032570  -7.970 2.19e-15 ***
cleaned_data$Age                 -0.008663   0.001076  -8.054 1.12e-15 ***
cleaned_data$EthnicAfrican-American -0.738348   0.054766 -13.482  < 2e-16 ***
cleaned_data$EthnicAsian         -0.530223   0.056770  -9.340  < 2e-16 ***
cleaned_data$EthnicHispanic      -0.738540   0.041412 -17.834  < 2e-16 ***
cleaned_data$EthnicOther         -0.224759   0.183252  -1.226   0.2201
cleaned_data$Health              -0.212388   0.015749 -13.486  < 2e-16 ***
cleaned_data$HeatBreaks          -0.042887   0.017648  -2.430   0.0151 *
cleaned_data$MiceRatsYes         -0.222819   0.044410  -5.017 5.53e-07 ***
cleaned_data$CracksHolesYes       0.045415   0.054812   0.829   0.4074
cleaned_data$BrokenPlasterYes    -0.042569   0.047525  -0.896   0.3705
cleaned_data$WaterLeakageYes     -0.030570   0.046452  -0.658   0.5105
cleaned_data$MaintenanceDef       0.004283   0.019893   0.215   0.8295
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9034 on 3205 degrees of freedom
Multiple R-squared:  0.3465,    Adjusted R-squared:  0.3438
F-statistic: 130.7 on 13 and 3205 DF,  p-value: < 2.2e-16
```

Figure 2:

```
Analysis of Variance Table

Model 1: log(cleaned_data$Income) ~ cleaned_data$Gender + cleaned_data$Age +
    cleaned_data$Ethnic + cleaned_data$Health + cleaned_data$HeatBreaks +
    cleaned_data$MiceRats + cleaned_data$WaterLeakage
Model 2: log(cleaned_data$Income) ~ cleaned_data$Gender + cleaned_data$Age +
    cleaned_data$Ethnic + cleaned_data$Health + cleaned_data$HeatBreaks +
    cleaned_data$MiceRats + cleaned_data$CracksHoles + cleaned_data$BrokenPlaster +
    cleaned_data$WaterLeakage + cleaned_data$MaintenanceDef
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   3208 2616.9
2   3205 2615.9  3   0.99553 0.4066 0.7483
```

Model 1:

```
first_model = lm(log(income) ~ age + gender + ethnic + health + heat_breaks + mice_rats + cracks_holes + broken_plaster +
water_leakage + maintenance_def)
```

Model 2 (cleaned_data removes the outliers and truncated data):

```
plot(x=cleaned_data$Age, y=cleaned_data$Income, main="Age v. Income", xlab="Age (years)", ylab="Income")
hist(cleaned_data$Income, main="Income", breaks=50)
```

Model 3:

```
model_wo_outliers = lm(log(cleaned_data$Income) ~ cleaned_data$Gender + cleaned_data$Age + cleaned_data$Ethnic +
cleaned_data$Health + cleaned_data$HeatBreaks + cleaned_data$MiceRats + cleaned_data$CracksHoles +
cleaned_data$BrokenPlaster + cleaned_data$WaterLeakage + cleaned_data$MaintenanceDef)
```

Model 4:

```
fit_model = lm(log(cleaned_data$Income) ~ cleaned_data$Gender + cleaned_data$Age + cleaned_data$Ethnic +
cleaned_data$Health + cleaned_data$HeatBreaks + cleaned_data$MiceRats)
```

Model 5-10:

```
fit_model_deg_2 = lm(log(cleaned_data$Income) ~ cleaned_data$Gender + poly(cleaned_data$Age, degree=2) +
cleaned_data$Ethnic + cleaned_data$Health + cleaned_data$HeatBreaks + cleaned_data$MiceRats)
fit_model_deg_3 = lm(log(cleaned_data$Income) ~ cleaned_data$Gender + poly(cleaned_data$Age, degree=3) +
cleaned_data$Ethnic + cleaned_data$Health + cleaned_data$HeatBreaks + cleaned_data$MiceRats)
fit_model_deg_4 = lm(log(cleaned_data$Income) ~ cleaned_data$Gender + poly(cleaned_data$Age, degree=4) +
cleaned_data$Ethnic + cleaned_data$Health + cleaned_data$HeatBreaks + cleaned_data$MiceRats)
fit_model_deg_5 = lm(log(cleaned_data$Income) ~ cleaned_data$Gender + poly(cleaned_data$Age, degree=5) +
cleaned_data$Ethnic + cleaned_data$Health + cleaned_data$HeatBreaks + cleaned_data$MiceRats)
fit_model_deg_6 = lm(log(cleaned_data$Income) ~ cleaned_data$Gender + poly(cleaned_data$Age, degree=6) +
cleaned_data$Ethnic + cleaned_data$Health + cleaned_data$HeatBreaks + cleaned_data$MiceRats)
fit_model_deg_7 = lm(log(cleaned_data$Income) ~ cleaned_data$Gender + poly(cleaned_data$Age, degree=7) +
cleaned_data$Ethnic + cleaned_data$Health + cleaned_data$HeatBreaks + cleaned_data$MiceRats)
```