Housing Condition in New York City

Steffi Chern (steffic)

November 22, 2022

Introduction

The New York City Housing and Vacancy Survey is conducted every three years to gain an insight into the recent housing situations in New York (especially the relationship between household income and several demographic and housing quality measurements). Understanding these information helps government make the necessary decisions and policies to improve the housing conditions. **(1)** In this report, we intend to answer the following research questions:
- whether or not there is a difference in average household income between Caucasians and Hispanics while controlling for other variables
- if there is a difference in the relationship between age and household income when water leakage has occurred in the apartment or not while controlling for other variables
For this study, we hypothesize that the Caucasian population would have a higher average household income than that of the Hispanic population when other predictors are held constant, since previous studies have shown such a pattern. In addition, we hypothesize that the relationship between age and the household income is not significantly different depending on whether or not there is water leakage in the apartment when other predictors are held constant, since it doesn't seem like the events have much relation with each other.

Exploratory Data Analysis

2.1 Data

(2) In the original dataset, there are a total of 3373 observations. However, we noticed that some of the observations contain missing or truncated data. To prevent making inaccurate inferences later, we removed data that has income values below zero or over 9999999. A total of 81 observations are eliminated from the dataset (2 observations have income < 0, and 79 observations have income > 9999999). There are 3292 observations remaining that we will use to analyze in later sections.

2.2 Univariate EDA

From figure 1, we can look at the distributions of each quantitative variable through histograms. **(2)** We observe that the Age variable is normally distributed. However, the variables Income, Heat Breaks, and Maintenance Deficiencies are all severely skewed to the left, which may raise concerns regarding high leverage points and nonlinearity. **(7)** To address these potential problems, we take the log transformation of the response variable Income. Note that we do not transform the predictors Heat Breaks and Maintenance Deficiencies because the transformations do not change much of their overall distribution.

(7) After taking the log transformation of Income, we notice from the transformed histogram that it is approximately normally distributed now. Thus, we will be moving forward with this transformed variable to do further analysis.

To understand the individual categorical variables, we can take a look at the summary table (not shown here). **(2)** We observe an imbalanced number of datapoints for the binary variables MiceRats, CracksHoles, BrokenPlaster, and WaterLeakage (a lot more No's than Yes's). This could be a concern when we answer the second research question regarding if there is a difference in the relationship between age and the household income when there is water leakage in the apartment or not. There are significantly more datapoints that are identified as Caucasians or Hispanics in the Ethnic variable, but this shouldn't be a concern when we answer the first research question since we will not take into account other ethnicities. There are slightly more males than females in this dataset (1851 vs 1441). For the Health variable, there are significantly more individuals who rated themselves in the 1-3 range than the 4-6 range.

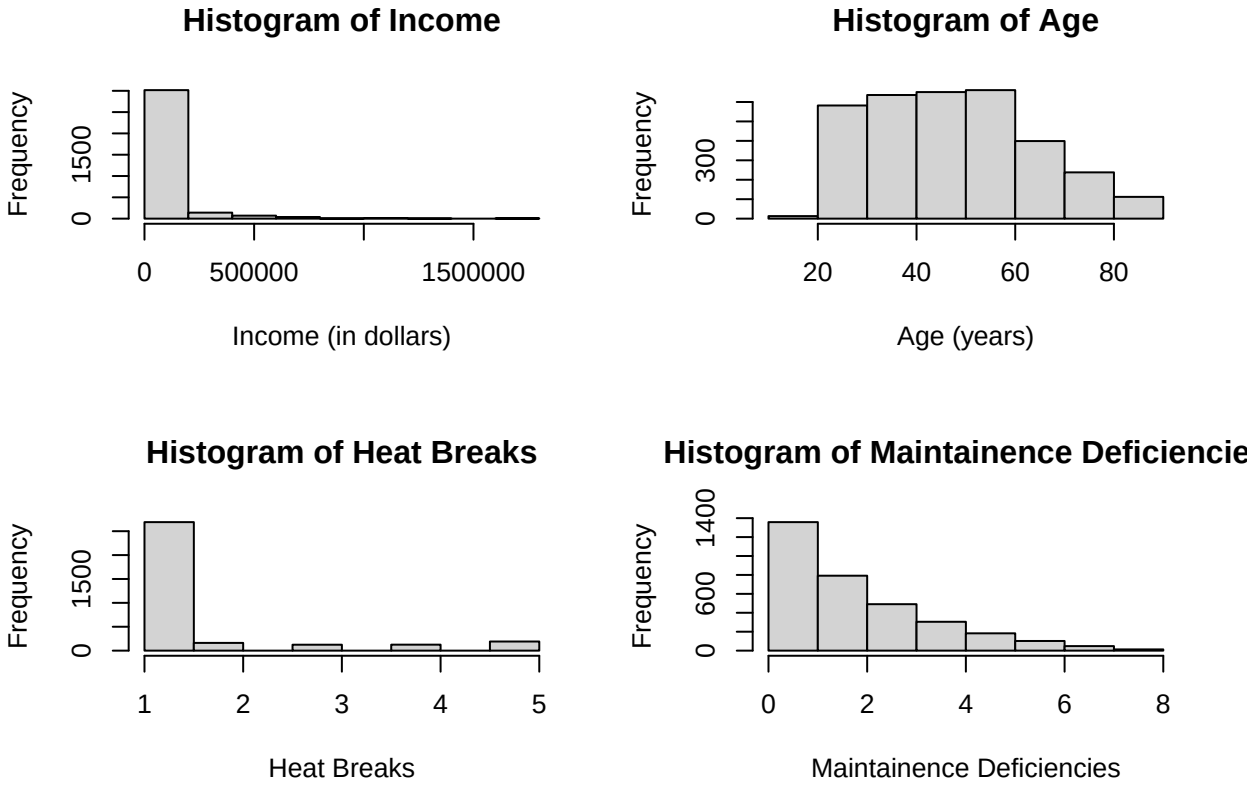


Figure 1: Histograms of Quantitative Predictors, Response Variable

Figure 2: Histogram of Transformed Response Variable

2.3 Bivariate EDA

(3) From figure 3 below, we see a weak negative relationship with Log of Income vs. Age and Log of Income vs. Maintenance Deficiencies. Moreover, there seems to be a positive relationship between the predictors Heat Breaks and Maintenance Deficiencies, but it doesn't seem to result multicollinearity problems since the relationship is relatively weak. The remaining predictors seem to be independent of each other.

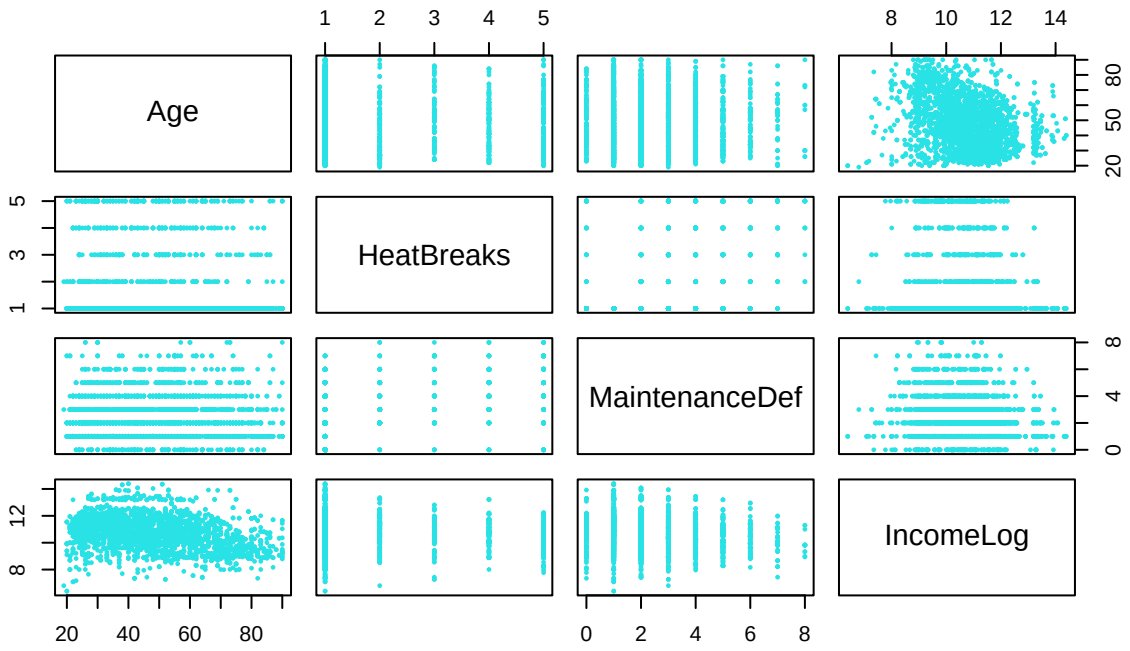


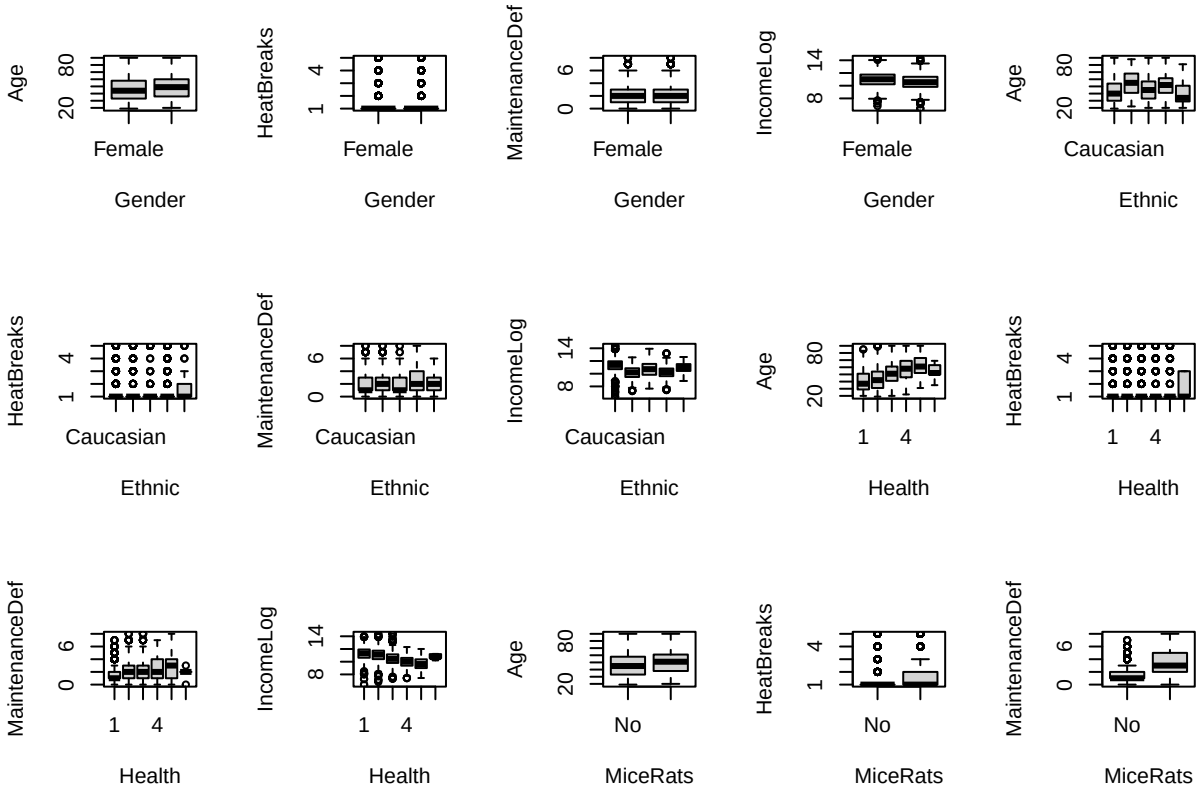
Figure 3: Pairs Plot of Quantitative Predictors and Response Variable

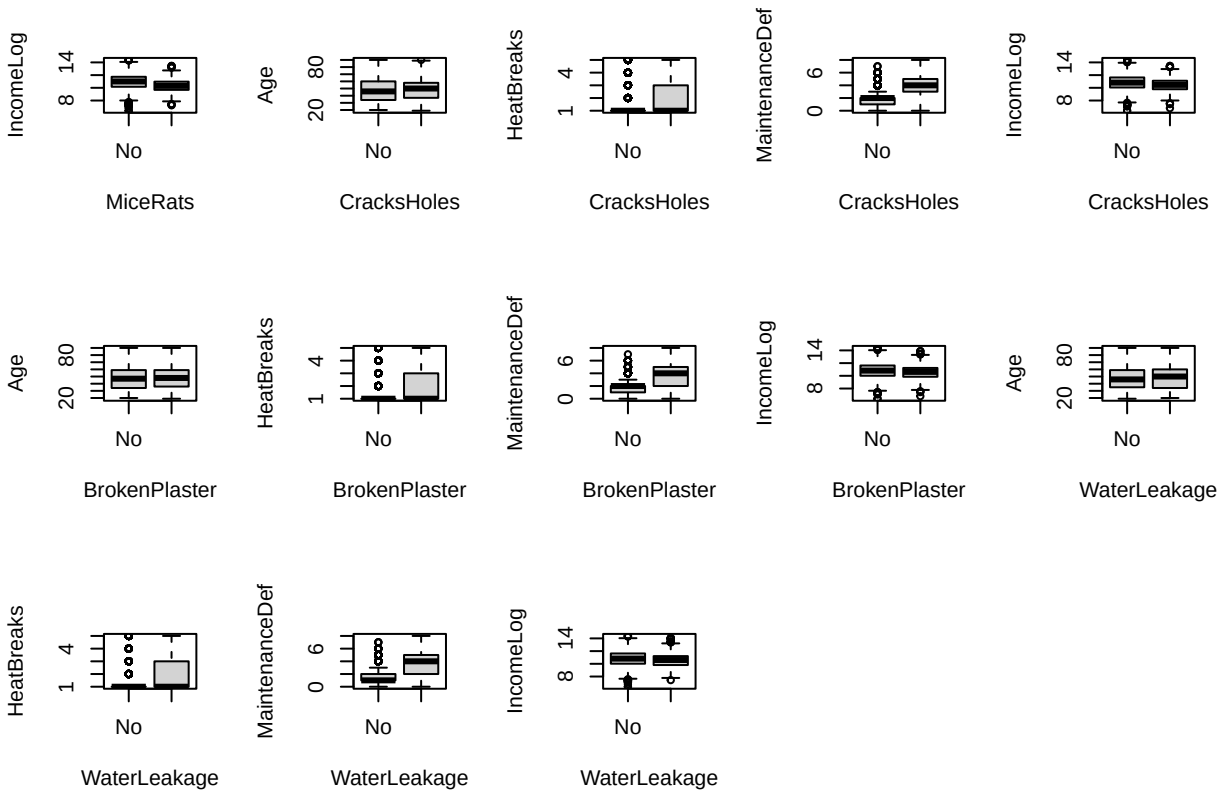
(3) Inspecting the relationship between quantitative and categorical variables, we notice a positive relationship between Age vs. Health. This seems reasonable since elders tend to have more illnesses or other health problems compared to youngsters. We also notice that the number of Maintenance Deficiencies is higher when there are Mice/Rats, Crack Holes, and Broken Plaster.

Also from looking at the boxplots below, we observe that Hispanics have lower median household income compared to that of Caucasians. Moreover, males have slightly lower median household income than that of females.

Inspecting the relationship between each categorical variable (table not shown here), we see that there are a lot more Caucasians who do not have MiceRats in their houses than

those who do. However, the number of Hispanics who do not have MiceRats in their houses are very close to the number of Hispanics who do.
Since there are no clear relationships between the other predictors themselves, thus they are most likely independent from each other.





Initial Modeling and Diagnostics

(4) As an initial model, we first fit a multiple linear regression model between the log Income (response variable) and all other variables as predictors.
We treat the variables Gender, Ethnic, MiceRats, CracksHoles, BrokenPlaster, and WaterLeakage as categorical variables, since they are either binary in nature or have values that can be categorized reasonably. In contrast, we treat the variables Log Income, Age, HeatBreaks, and MaintenanceDef as continuous variables, since they can all take on either real values or positive integer values.
We note that Health is treated as an ordinal variable, since it takes on numbers 1 to 6 that indicates the level of health status (higher means better health status).

(6) We calculated the Cook's distance to check for unduly influential observations before any interactions are added. The three most influential points (highest Cook's distance) have been identified graphically. We compared the distances to the quantiles of the F distribution with 18 and n-18 (3274 in this case) degrees of freedom. **(8)** Since all observa-

tions did not exceed the 50th percentile of the F distribution (0.96), we conclude that none of the observations are overly influential, thus there doesn't seem to exist outliers in this case.

(6) When we inspect the residuals vs fitted values plot, we notice the variance of the residuals are mostly constant (residuals evenly spread above and below the red zero line), but slightly increases when the fitted values increases. The residuals are patternlessly scattered above and below the horizontal zero line. In addition, the conditional expectation of errors given the predictors is approximately 0.

(6) Based on the Normal Probability Plot, we can assess whether the assumption that irreducible errors are normally distributed is validated or not. The majority of the residuals still lie along the diagonal line. However, we observe the lower tail of Q-Q plot curving a bit away from the diagonal line, which aligns with what we observe about the residuals vs fitted values plot. Overall, the model assumptions are not satisfied entirely, but this shouldn't be too much of a concern for us when making inferences later.

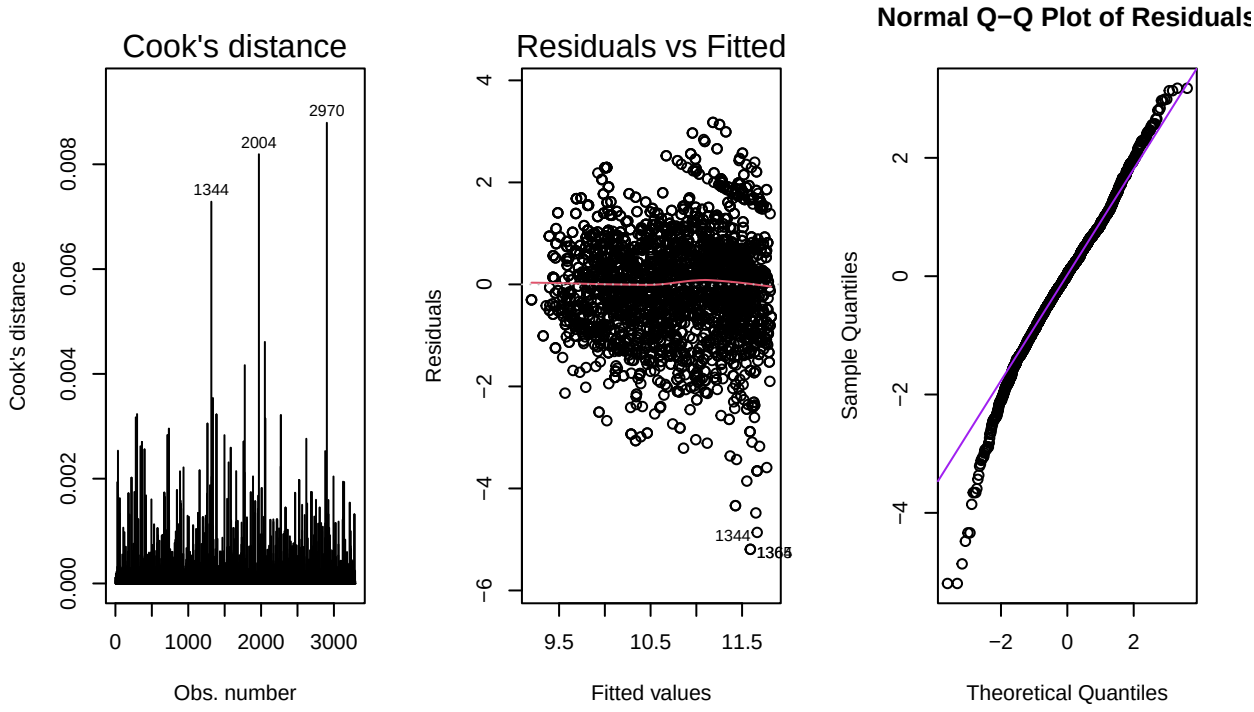


Figure 4: Diagnostic Plots Without Interactions

(6 and 7) To better fit the model assumptions, we could add some meaningful interactions. In general, transformations of variables may work too, but we already transformed our response variable and we discovered that other variables do not show much distri-

butional improvement after transformations.

(5) To address the second hypotheses mentioned in the introduction section, we add the interaction Age:WaterLeakage. **(4 and 6)** As a possible step to improve our model, we could try adding the interaction Gender:Ethnic since we previously observed that there is a possible difference in the relationship between Gender and Income, and this may be depended on which ethnicity it is.

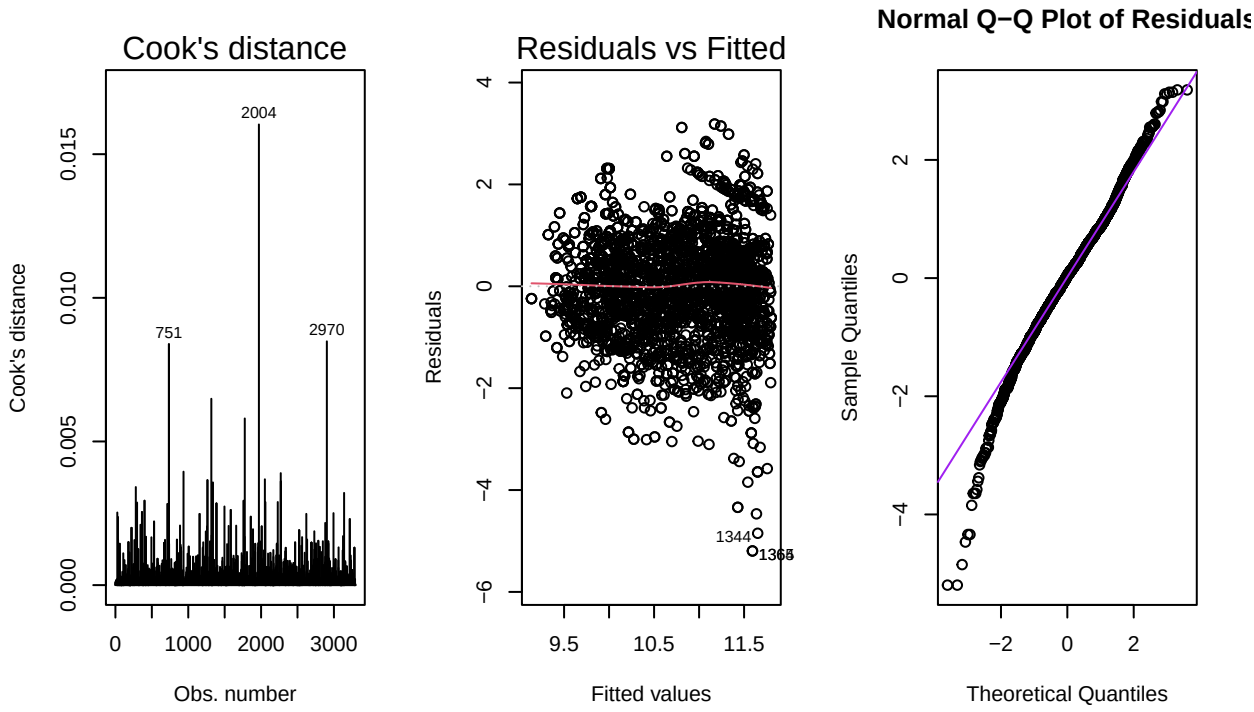


Figure 5: Diagnostic Plots With Interactions

After taking into account the interaction terms, we see that the diagnostic plots look nearly identical to the ones without the interaction terms. **(8)** The three most influential points in this case have higher cook's distance in general than the ones in the previous case, but there doesn't seem to exist any outliers since the distances are all lower than the 50th percentile of the F distribution with 23 and n-23 (3269 in this case) degrees of freedom.

(9) Since our adjustments to the model did not help with fitting the model assumptions better, we will just proceed with the model without interactions (our initial model). Despite the model assumptions not fully satisfied, the model is well enough for us to make reasonable interpretations.

Model Inference and Results

(10) To answer the first research question regarding whether or not there is a difference in the average household income for Caucasians and Hispanics when other variables are held constant, we proceed with using the initial model (no interaction terms) and conduct the following hypothesis test:
$H_0 : \beta_{Hispanic} = 0$
$H_a : \beta_{Hispanic} \neq 0$
Since we manually set Caucasian as the baseline level for the Ethnic variable, we can proceed with just testing a single beta coefficient.

When we test at the 95% confidence level, we get a critical value around 1.96. According to the summary (not shown here), we observe that the t-value associated with EthnicHispanic is -15.139, and the absolute value of it is greater than 1.96. Thus, we reject the null hypothesis that $\beta_{Hispanic}$=0. We can conclude that there is a significant difference in the average household income for the Caucasian population and the Hispanic population when various demographic and housing quality measurements are held constant. We can also verify this by inspecting the p-value. The p-value associated with EthnicHispanic is 2e-16, which is less than 0.05, confirming that we should reject the null hypothesis.

To answer the second research question about whether or not the relationship between Age and the Household Income is different depending on the occurrence or nonoccurrence of Water Leakage (while other variables are controlled), we conduct the following hypothesis test:
$H_0 : \beta_{Age:WaterLeakage} = 0$
$H_a : \beta_{Age:WaterLeakage} \neq 0$

The first model has no interaction terms, while the second model includes the Age:WaterLeakage interaction. We compare the two models by performing an ANOVA test (not shown here). We get a test statistics (F-statistic) of 4e-04. We compare the F-statistic to the F distribution with 1 and 3273 degrees of freedom. Since the F-statistic is less than the critical value of the F-distribution (4e-04 < 5.0285), we fail to reject the null hypothesis. We can conclude that there is no sufficient evidence that the relationship between Age and Household Income is different depending on whether Water Leakage occurred or not in the apartment, while controlling for other variables. We can also verify

this by inspecting the p-value. Since the p-value is 0.9842, which is greater than 0.05, this confirms that we should not reject the null hypothesis.

(11) To create a model with the best prediction performance, we performed model selection using bestglm() along with AIC as the statistically rigorous criterion to minimize the prediction error.
The following predictors are included in the final model: Gender, Age, all Ethnic levels, all Health levels, HeatBreaks,and MiceRats. There are now 13 predictors instead of 17 predictors.

Conclusion and Discussion

Understanding the current housing situations is crucial in improving our living standards. **(12)** After thorough analyses of this given dataset, we can conclude two critical findings: There is sufficient evidence to suggest that the average household income is different for Caucasians and Hispanics while holding other variables constant. This seems reasonable since looking back at the past history, there have been multiple instances where Caucasians are provided with much better opportunities (education, employment, etc.) for them to thrive in the future compared to Hispanics. Moreover, there is no sufficient evidence to suggest 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 characteristics are all the same. Intuitively speaking, water leakage should not have any influence on household income for different ages.
It is important to be aware that there are other predictors that weren't taken into consideration this time. In addition, the models generated in this study may not generalize well to households that have exceptionally low or high income, since we removed those observations before we started to utilize the dataset for further analysis. The imbalanced number of observations in the different levels of some categorical variables could potentially be a concern as well.