A Study of Data from the 2019 American Housing Survey

By Avneesh Verma

Introduction

This report is an analysis of data from the 2019 American Housing Survey, conducted by the U.S. Census Bureau. This is a longitudinal survey that studies the same set of housing units every other year. We hypothesize that there is a positive correlation between the market value of a family's house and the family's annual income. This is a large dataset that with over 60,000 rows (one for each household) and 1,000 columns (fields for each household). This analysis focuses on a few variables of interest: Family Income, Market Value, Monthly Housing Costs, and Unit Size.

Cleaning up data

The quantitative variables to be investigated are Family Income (Past 12 Months), Current Market Value of Unit, and Monthly Total Housing Costs (which is the sum of variables like Mortgage, Rent, Insurance, and Maintenance amounts per month). In this dataset, the number -6 represents "Not Applicable". Thus, a copy of the data will be created, and all the rows that have -6 for Income, Market Value, and Monthly Housing Costs will be removed from this copy.

```
* 8730 households have Family Income listed as 'Not Applicable', which is 13.8% of the total number of h

* 24795 households have Market Value listed as 'Not Applicable', which is 39.2% of the total number of h

* 8730 households have Monthly Housing Costs as 'Not Applicable', which is 13.8% of the total number of

* In total, there are 32971 households remaining in this filtered copy of the dataset. The original data
```

One Variable Analysis

Income

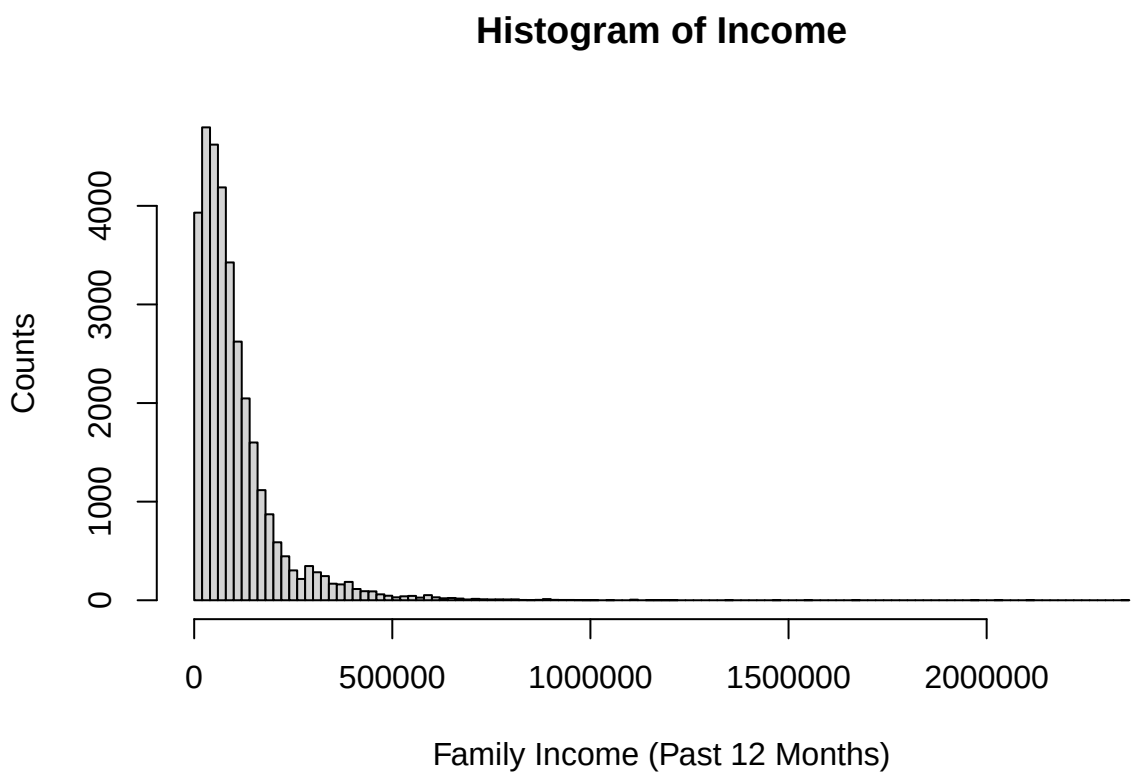
First, a histogram of "Family Income (Past 12 Months)" will be studied. This is the sum of all types of income for the householder and all related individuals age 16 and over.

Caution: We have also decided to remove **all** households whose income values are less than 0 (not just the ones that have -6). This is because it is not explained how households have negative income, and there is no way to distinguish between households that have an actual income of -$6 and households for which income is "Not Applicable". This is not a serious omission of data since there are very few households whose incomes are listed as negative numbers that are not -6:

```
Number of households with negative income (not including -6): 4
```

Now let's take a look at the histogram:

```r
options(scipen=5) # To avoid scientific notation x ticks
hist(
  income, main="Histogram of Income",
  xlab="Family Income (Past 12 Months)", ylab = "Counts", breaks=150
  )
```

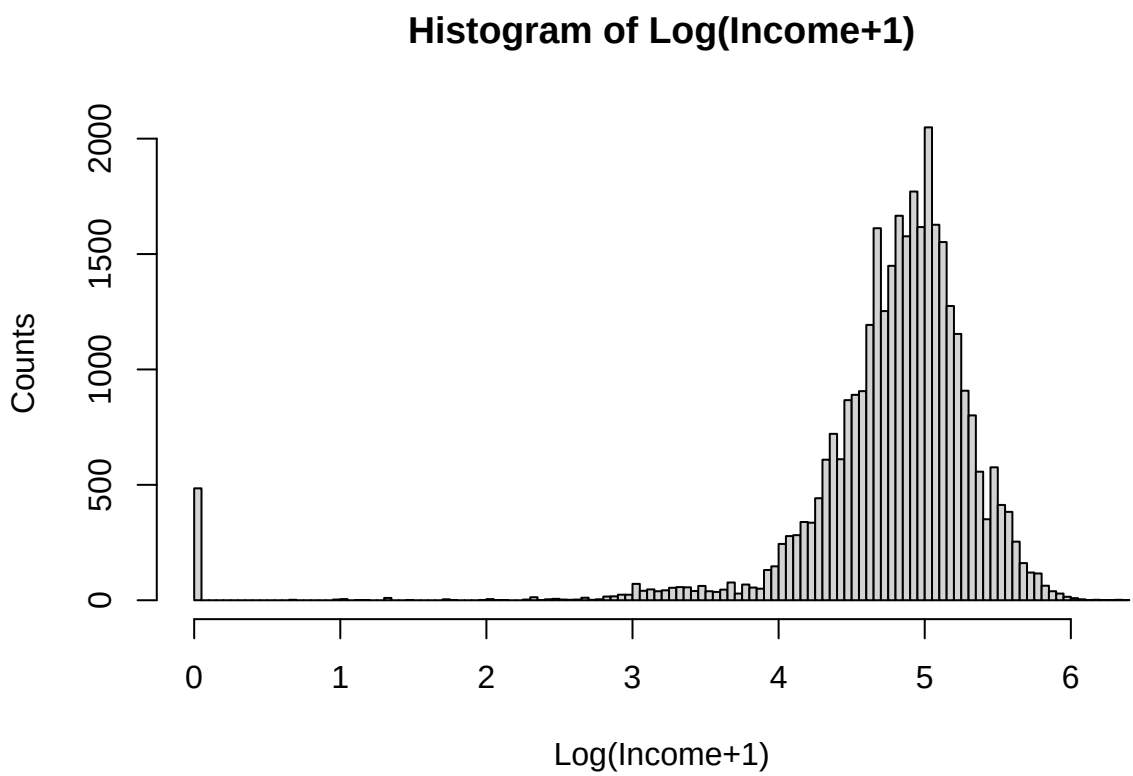



```
The mean Family Income is $104689.57
```

```
The standard deviation of Family Income is $109156.69
```

Clearly, the data is highly skewed to the right (which is not surprising for a variable like income). This results in a standard deviation that is greater than the mean. We will re-express the data by looking at a histogram of the log of income (base 10). Since we cannot take the logarithm of income values of "0", we'll first add a constant of 1 to each income value before taking the logarithm. This re-expressed data looks much more normal:

```
log_income <- log(income+1, base=10)
hist(
  log_income, breaks=150, main="Histogram of Log(Income+1)",
  ylab = "Counts", xlab = "Log(Income+1)"
)
```

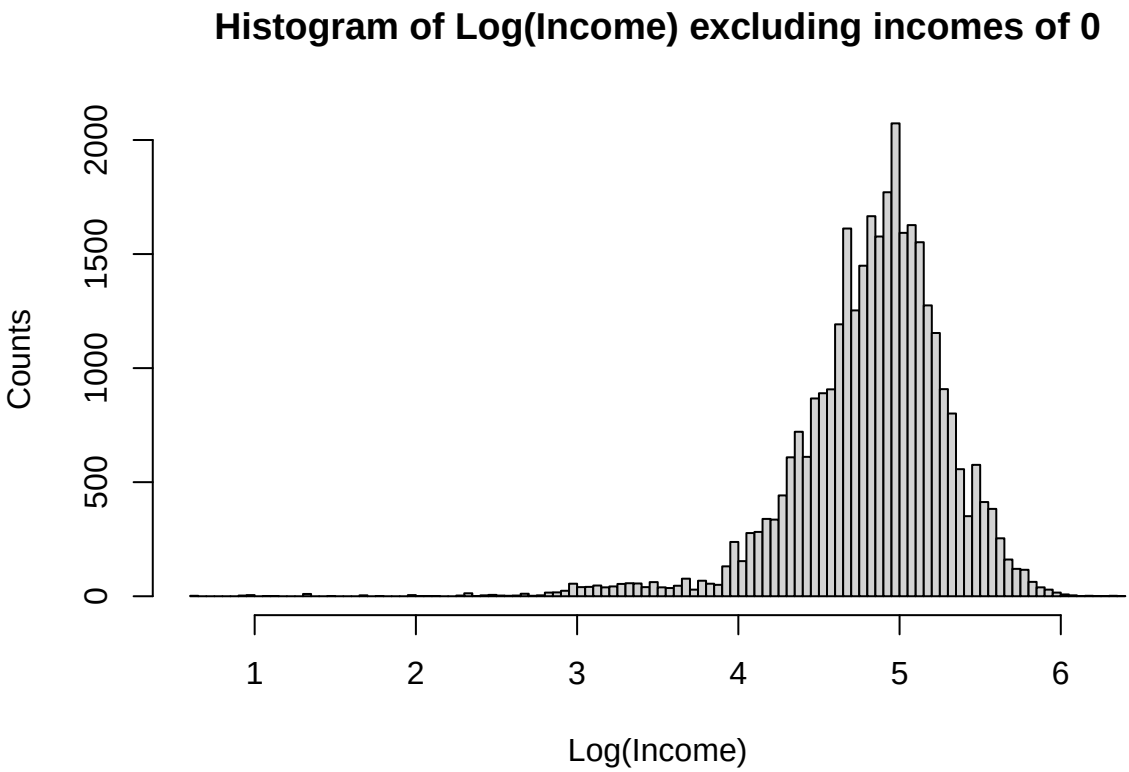



```
The mean of Log(Family Income+1) is 4.76
```

```
The standard deviation of Log(Family Income+1) is 0.748
```

There is a big spike at "0", which represents all of the households who reported a family income of $0 over the past 12 months. If we remove this outlier, the histogram looks and like this:

```
data_positive_income <- data[!(data$FINCP <= 0),]
log_income_no_zero = log(data_positive_income$FINCP, base=10)
hist(
  log_income_no_zero, breaks=150, main="Histogram of Log(Income) excluding incomes of 0",
  ylab = "Counts", xlab = "Log(Income)"
)
```




```
The mean of Log(Family Income) excluding incomes of $0 is 4.831
```

```
The standard deviation of Log(Family Income) excluding incomes of $0 is 0.474
```

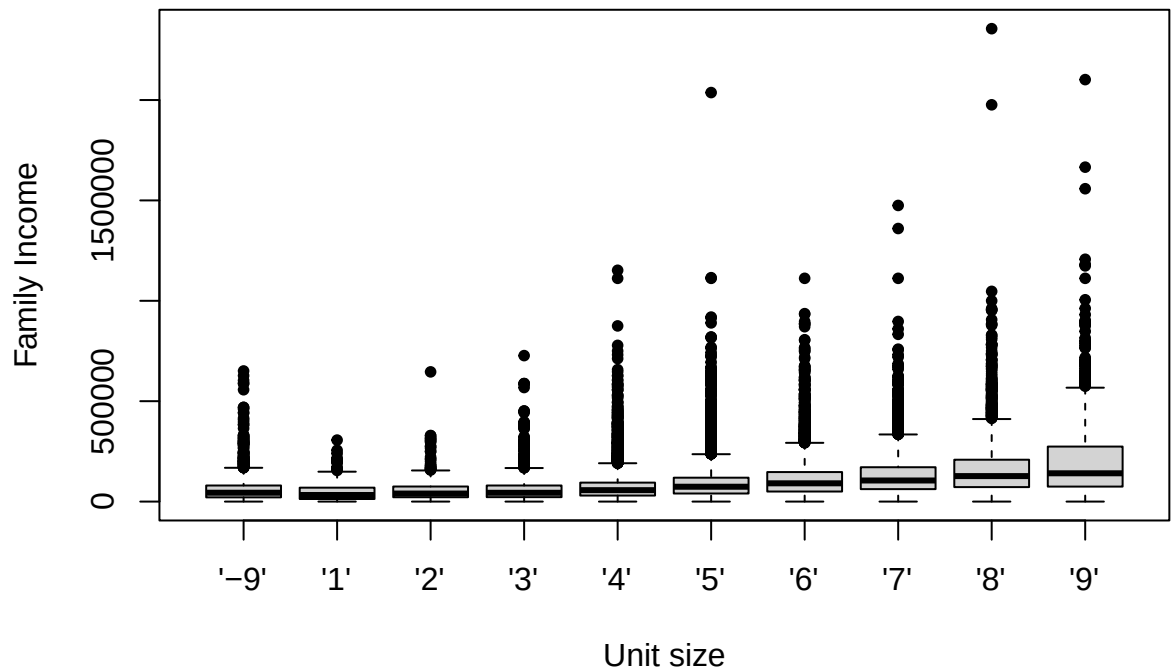
This outlier spike at 0 had pulled the mean to the low end, so removing these outliers results in a larger mean and a smaller standard deviation.

Let's take a look at the relationship between income and house size. In this dataset, unit size (in square footage) is reported as a categorical variable, with each category representing an interval in square feet:

- '-9': Not Reported
- '1': Less than 500 square feet
- '2': 500 to 749 square feet
- '3': 750 to 999 square feet
- '4': 1,000 to 1,499 square feet
- '5': 1,500 to 1,999 square feet
- '6': 2,000 to 2,499 square feet
- '7': 2,500 to 2,999 square feet
- '8': 3,000 to 3,999 square feet
- '9': 4,000 square feet or more

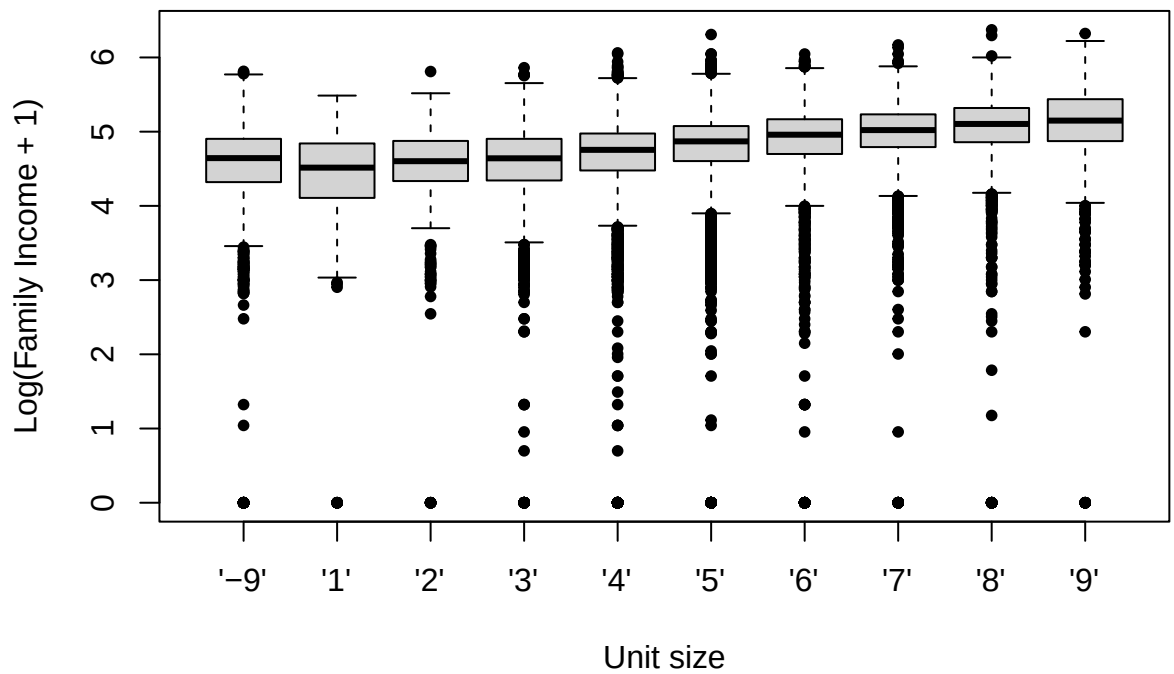
We hypothesize that there is a positive association between unit size and family income. It seems reasonable that families that make more money would purchase larger houses.

```
boxplot(income ~ data$UNITSIZE, xlab = "Unit size", ylab = "Family Income", pch=20)
```



This is hard to interpret, so we'll plot Log(Income+1) against unit size instead.

```
bp <- boxplot(
  log_income ~ data$UNITSIZE,
  xlab = "Unit size", ylab = "Log(Family Income + 1)", pch=20
)
```




```
Medians:
4.643463 4.515212 4.602071 4.640988 4.754738 4.869238 4.959046 5.021193 5.103807 5.148914
Interquartile Ranges:
```

```
0.5829284 0.7338563 0.5405932 0.560653 0.4978409 0.4718546 0.4683416 0.4406 0.461021 0.5650075
```

Notice: as square footage increases, the log of median family income also tends to increase (excluding the -9 bracket, since that represents all households for which unit size is not reported.). Thus, our hypothesis appears to be correct.

One surprising result is that as square footage increases, the interquartile range of Log(Family Income + 1) tends to decrease. One exception to this rule is the '9' bracket, since it represents all households which have a unit size greater than 4000 square feet, while the other brackets have a fixed upper limit of square feet. Bracket '8' is also an exception, because the range of square footage in this bracket is 1000 square feet, while the ranges of square footage in the most of the other brackets are 500 square feet. Bracket '3' is also an exception, for unknown reasons.

One possible explanation for this negative association between the interquartile ranges and unit sizes is that the smaller spread for larger unit sizes is a result of the cost barrier blocking people from purchasing these houses. In other words, a family that makes a lot of money may still *choose* to live in a home of modest size (resulting in a greater spread in incomes for smaller house sizes). However, a family that makes a small amount of money simply *cannot afford* to buy and live in a bigger house, resulting in the market for larger homes being dominated solely by wealthier families.

Market Value

Next, let's study the variable respresenting "Current Market Value of Unit". First, let's look at the five number summary.

```
Min: 1000

Q1: 149002

Median: 262086

Q3: 440042

Max: 9999998
```

In our dataset, the number 9,999,998 actually represents all house prices including *or greater than* that number. Thus, we will remove all households with values of 9,999,998 since it is impossible to know their actual price.

```
21 households have Monthly Housing Costs listed as '9999998', which rounds down to 0% of the total numbe
```

Since there are only 21 houses with market values listed as "9,999,998", removing them is not a significant omission. Here's the new five number summary:

```
Min: 1000

Q1: 148991.5

Median: 262054

Q3: 439911

Max: 9991165
```

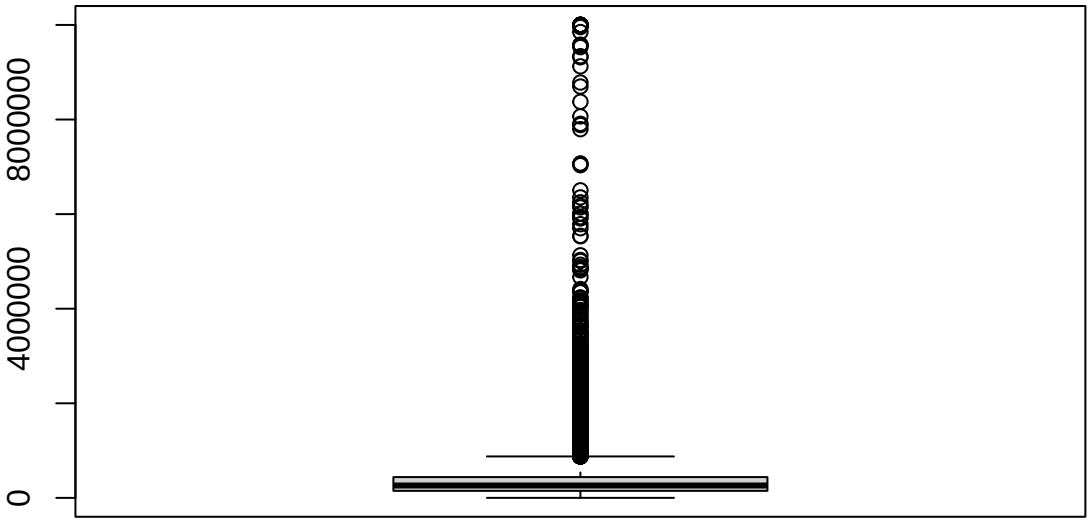
The five number summary reveals a few key facts:

- There are no houses with a market value less than $1000 in the dataset. Perhaps those conducting the survey did not consider houses cheaper than $1000 to be "valid" homes that could be put on the market.
- About 25% of houses have a market value less than $148,991.50, and about 25% of houses have a market value greater than $439,911.00. The middle 50% of houses have market values between those two values.
- About 50% of houses have market values below $262,054.00, and 50% of houses have market values above that number.

- The most expensive house remaining in the dataset (after all of the aforementioned filters) has a market value of $9,991,165.00.
- The median in this five number summary ($262,054), in which the extreme values on the high end had been removed, is only slightly lower than the median in the previous five number summary ($262,086). This is an example of how medians are resistant to the pull of extreme values.

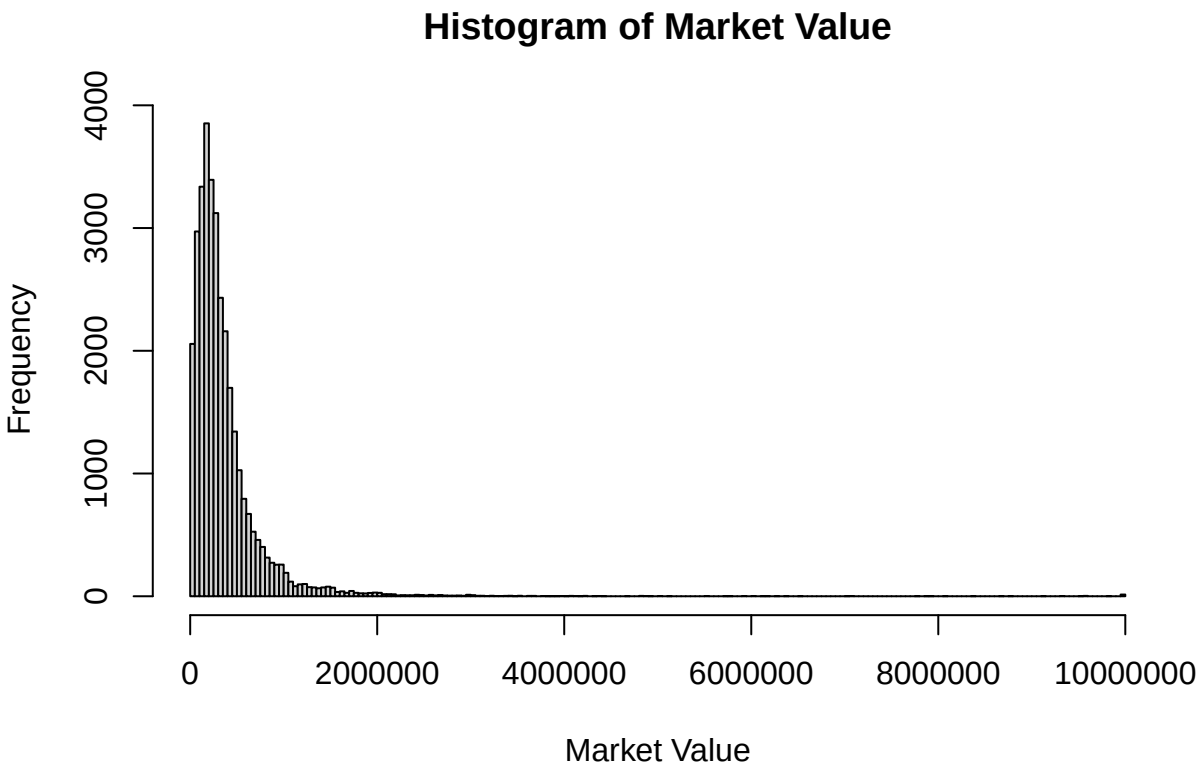
Now, let's look at the boxplot:

```
boxplot(market_val)
```



This boxplot is not very helpful, so let's look at the histogram.

```
hist(
  market_val, breaks=150, main='Histogram of Market Value', xlab='Market Value'
)
```

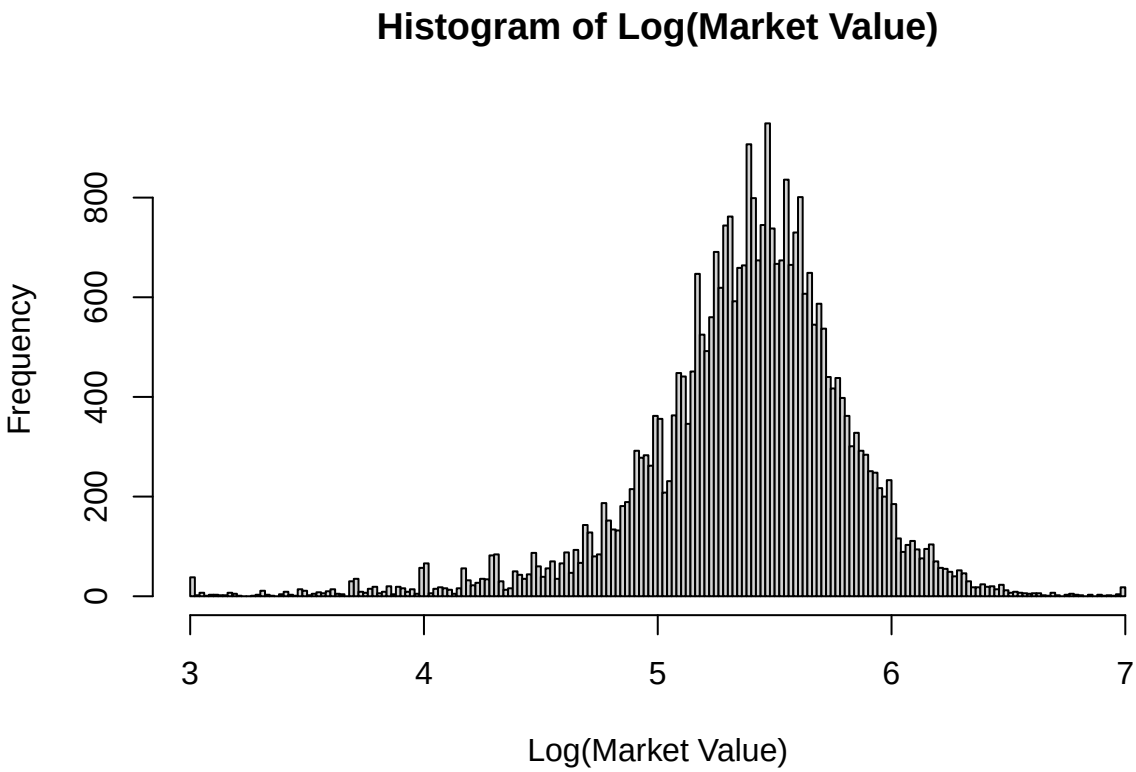



```
The mean of Market Value is $371066.09
```

```
The standard deviation of Market Value is 478495.873
```

Like Income, this is skewed to the right, so we'll re-express it by taking the logarithm (base 10). We don't need to worry about zeroes since the minimum is $1000.

```
log_market_val <- log(market_val, base=10)
hist(
  log_market_val, breaks=150,
  main="Histogram of Log(Market Value)", xlab = "Log(Market Value)"
)
```




```
The mean of Log(Market Value) is 5.377
```

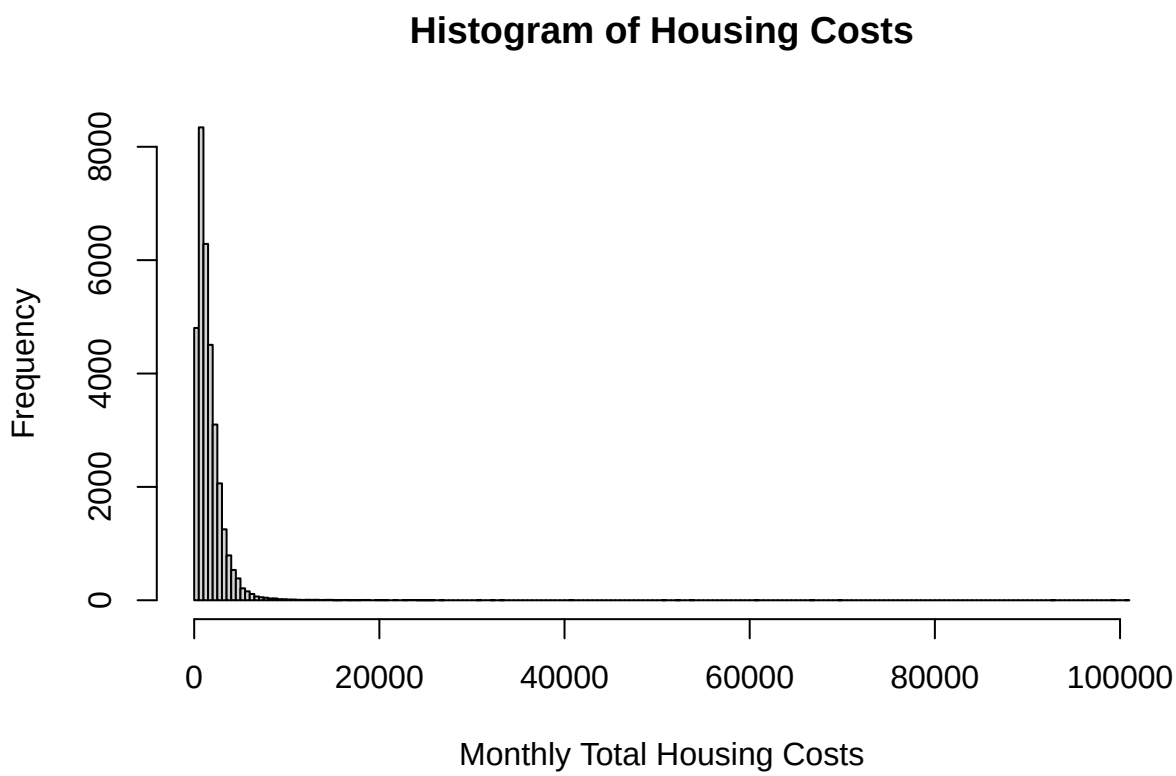
```
The standard deviation of Log(Market Value) is 0.446
```

Again, this looks much more normal. Notice that the x-axis begins at 3 because the minimum market value was 1000. It ends at 7 because the maximum was very close to $10,000,000. There are little peaks at the extremes. One possible explanation for the peak at 3 is that many houses with values around or below $1000 may have had their values rounded up to $1000. The peak at 7 appears for unkown reasons.

Housing Costs

Finally, let's look at "Monthly Total Housing Costs". This is the sum of variables like rent, mortgage, monthly utility bill, monthly insurance, monthly maintenance expense, etc.

```
housing_costs <- data$TOTHCAMT
hist(
  data$TOTHCAMT, breaks=150,
  main = "Histogram of Housing Costs", xlab = "Monthly Total Housing Costs"
)
```

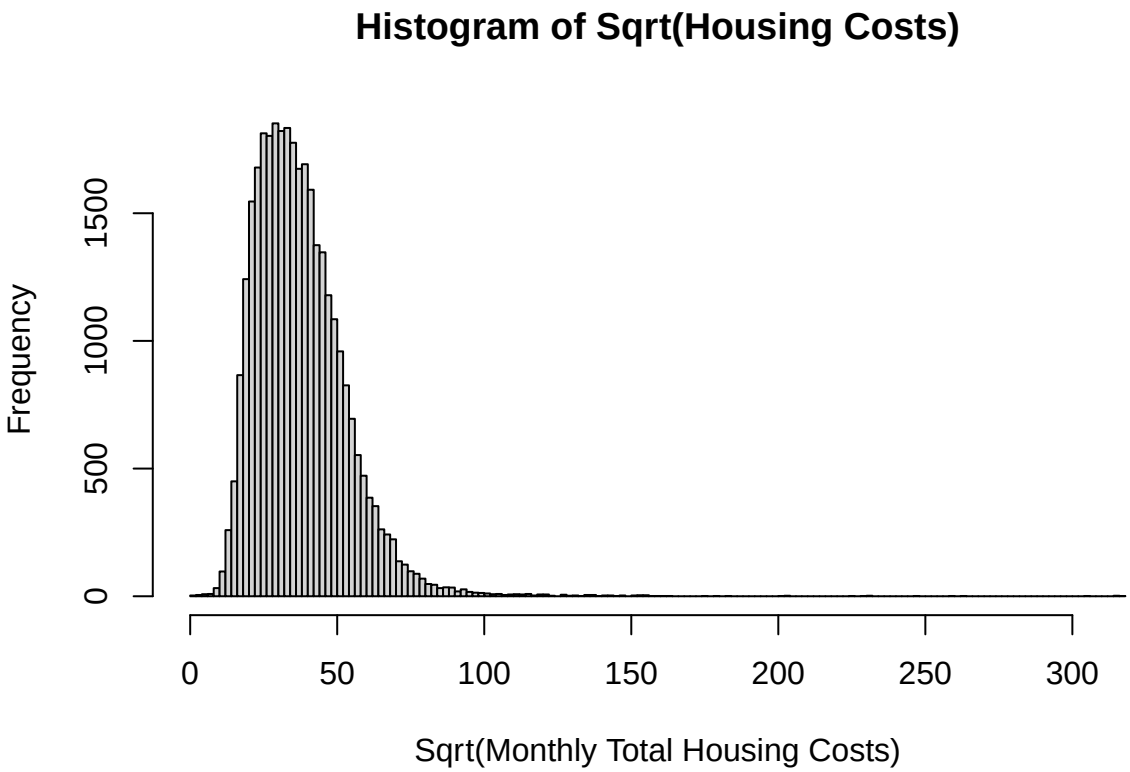



```
The mean of Housing Costs is $1665.29
```

```
The standard deviation of Housing Costs is 2103.31203878608
```

Here are some attempts of re-expression.

```
sqrt_housing_costs = sqrt(housing_costs)
hist(
  sqrt_housing_costs, breaks=150,
  main="Histogram of Sqrt(Housing Costs)", xlab="Sqrt(Monthly Total Housing Costs)"
)
```

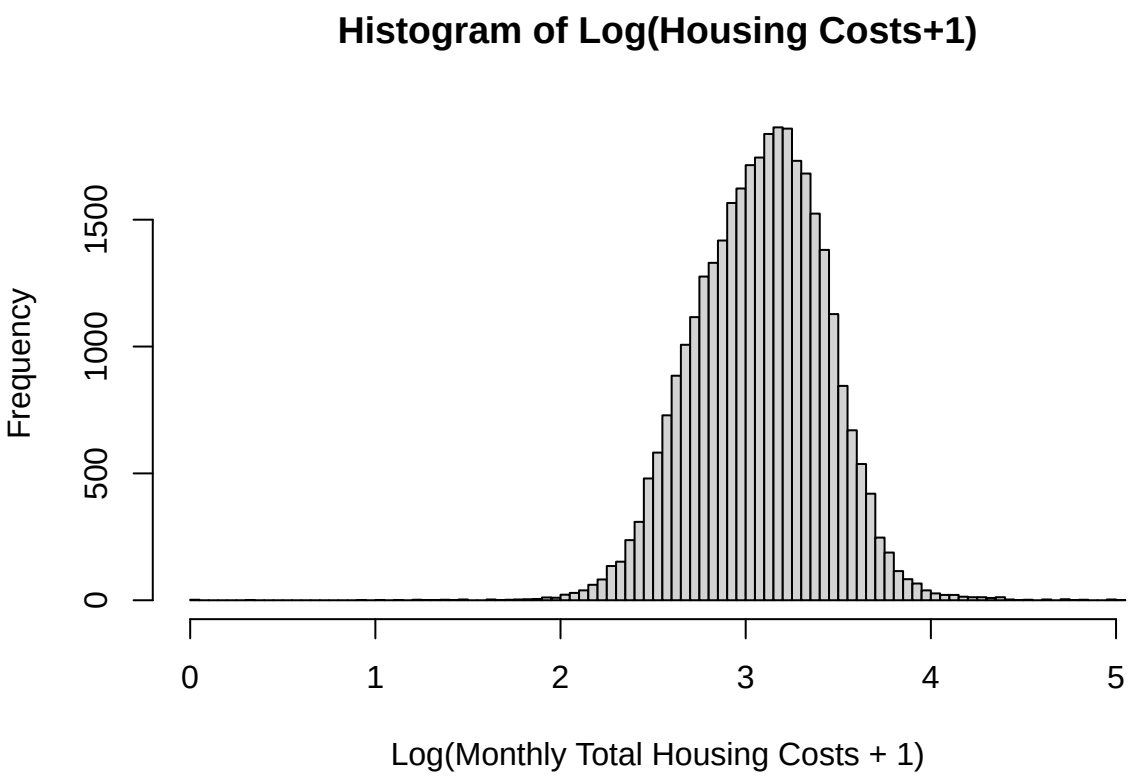



```
The mean of Sqrt(Housing Costs) is 37.551
```

```
The standard deviation of Sqrt(Housing Costs) is 15.976
```

The histogram of $\sqrt{\text{Housing Costs}}$ is still skewed to the right, so let's try taking the logarithm.

```
log_housing_costs = log(housing_costs+1, 10)
hist(
  log_housing_costs, breaks=150,
  main="Histogram of Log(Housing Costs+1)", xlab="Log(Monthly Total Housing Costs + 1)"
)
```




```
The mean of Log(Housing Costs + 1) is 3.08
```

```
The standard deviation of Log(Housing Costs + 1) is 0.35
```

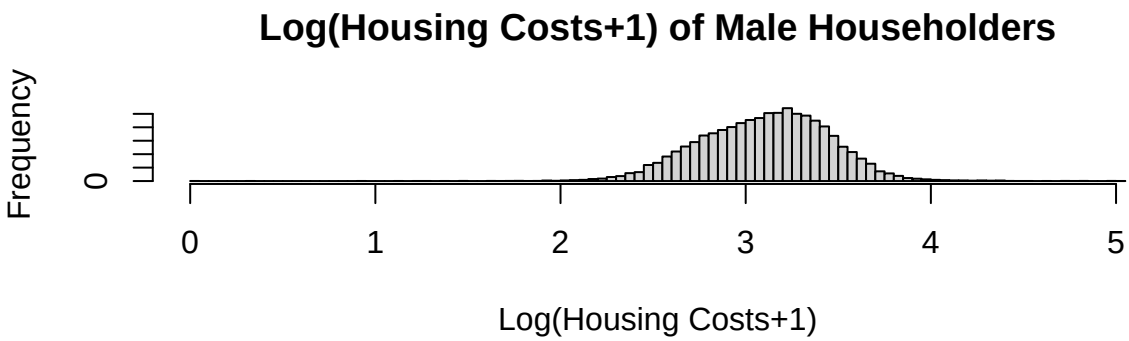
This works better.

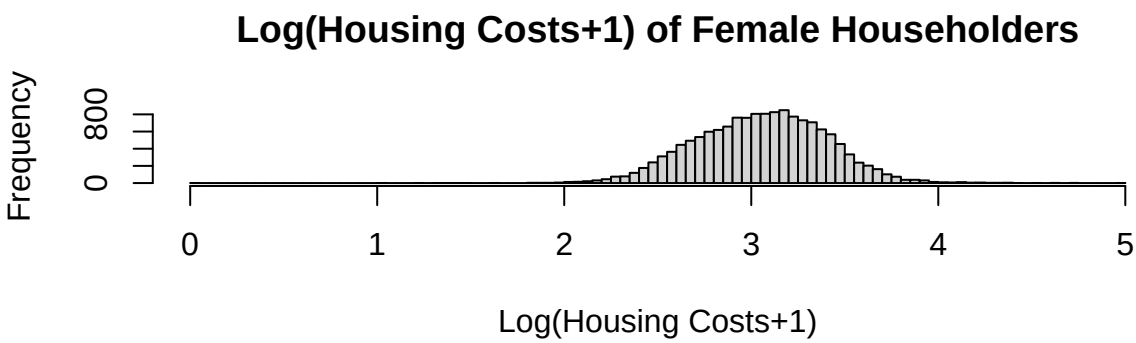
Bonus: Housing costs based on sex of householder.

While exploring our data, we found surprising relationship between total monthly housing costs and the sex of the householder. The definition of "householder" for this dataset is not clear, but we think it refers to the person being interviewed for the survey on behalf of the household.

```
sex <- data$HHSEX
housing_costs_male <- c()
housing_costs_female <- c()
# Store housing costs based on sex
for (i in 1:length(sex)) {
  if (sex[i] == "'1'") {
    housing_costs_male <- c(housing_costs_male, housing_costs[i])
  }
  else if (sex[i] == "'2'") {
    housing_costs_female <- c(housing_costs_female, housing_costs[i])
  }
}

log_housing_costs_male <- log(housing_costs_male+1, 10)
log_housing_costs_female <- log(housing_costs_female+1, 10)

par(mfrow=c(2,1)) # Display histograms side by side

hist(log_housing_costs_male, breaks = 150,
```

```
     main = "Log(Housing Costs+1) of Male Householders", xlab="Log(Housing Costs+1)")
hist(log_housing_costs_female, breaks=150,
     main = "Log(Housing Costs+1) of Female Householders", xlab="Log(Housing Costs+1)")
```






```
The mean Log(Housing Costs) for male householders is $3.1.

The standard deviation of Log(Housing Costs) for male householders is 0.349.

The mean Log(Housing Costs) for female householders is $3.05.

The standard deviation of Log(Housing Costs) for female householders is 0.349.
```

The histogram of Log(Housing Costs) female householders appears *very* slightly shifted to the left in comparison to the histogram for male householders. The fact that the mean Log(Housing Cost) for female householders is 5 cents less than the mean for male householders supports this. Of course, this difference is very small. Another point worth noting is that the standard deviations of Log(Housing Costs) are the same for both male and female householders; in other words, the spread is similar for both sexes.

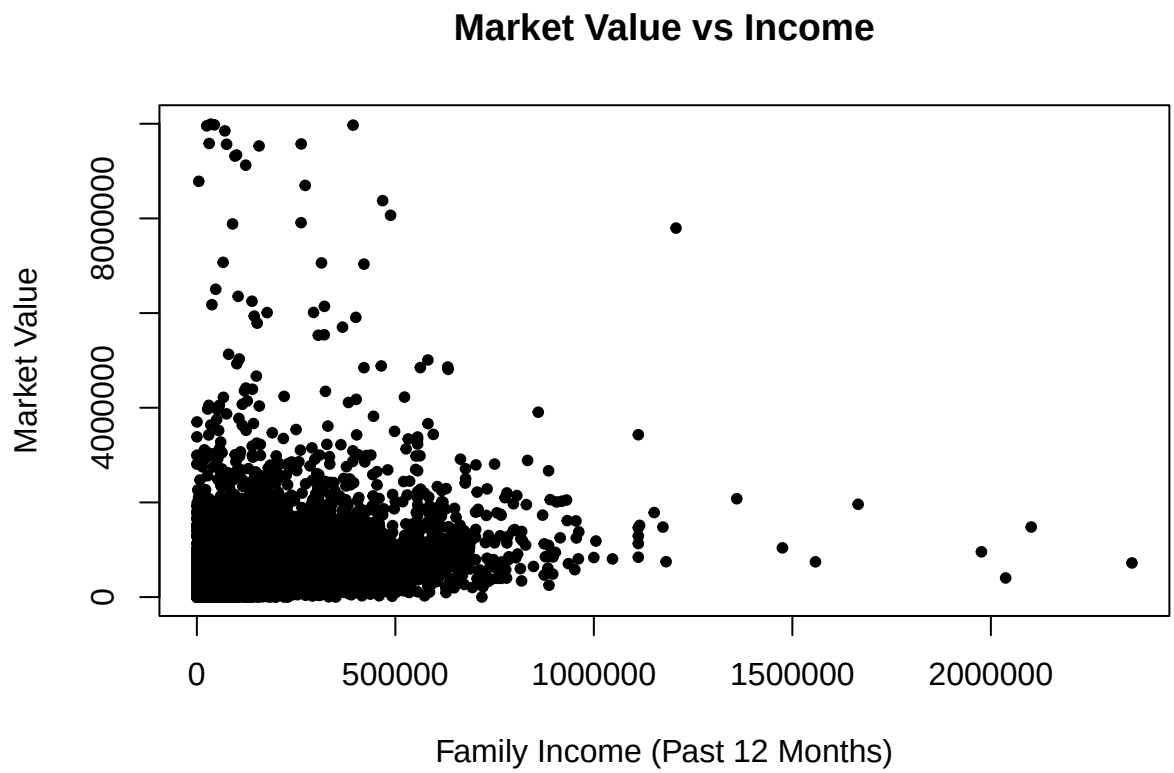
Two Variable Analysis

Recall: our hypothesis was that there is a positive correlation between the market value of a family's house and the family's annual income.

First, let's compare market value and income.

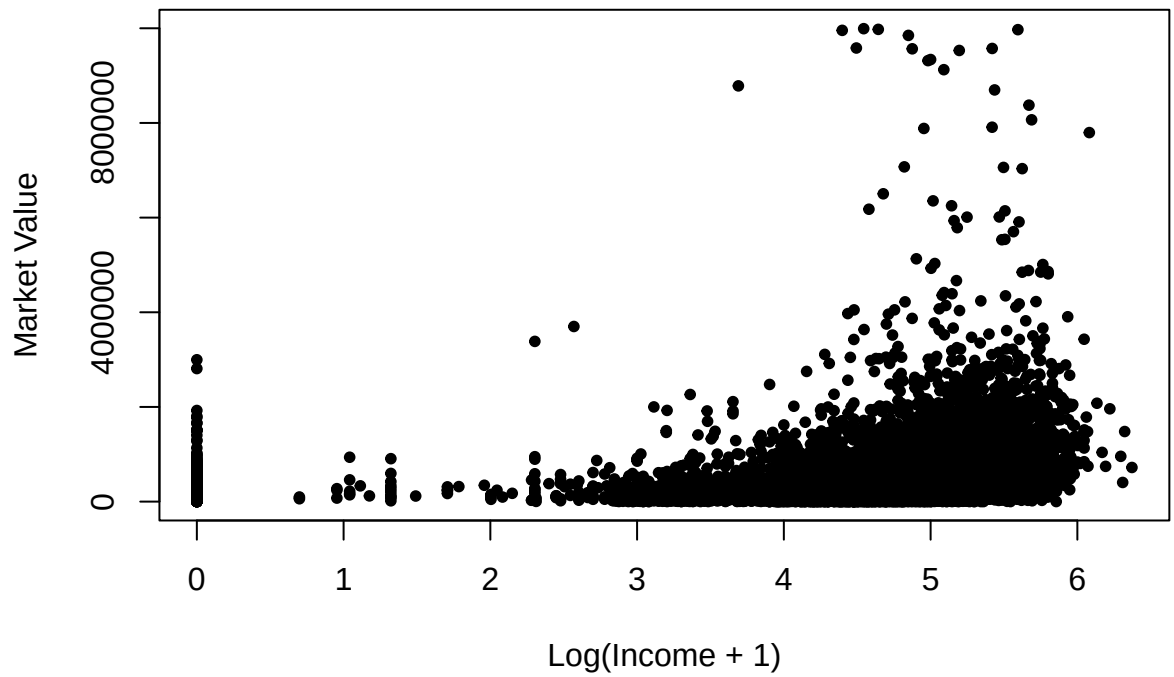
```
income <- data$FINCP
market_val <- data$MARKETVAL
plot(
  income, market_val,
  main="Market Value vs Income",
  xlab="Family Income (Past 12 Months)",
```

```
  ylab = "Market Value",
  pch = 20
)
```

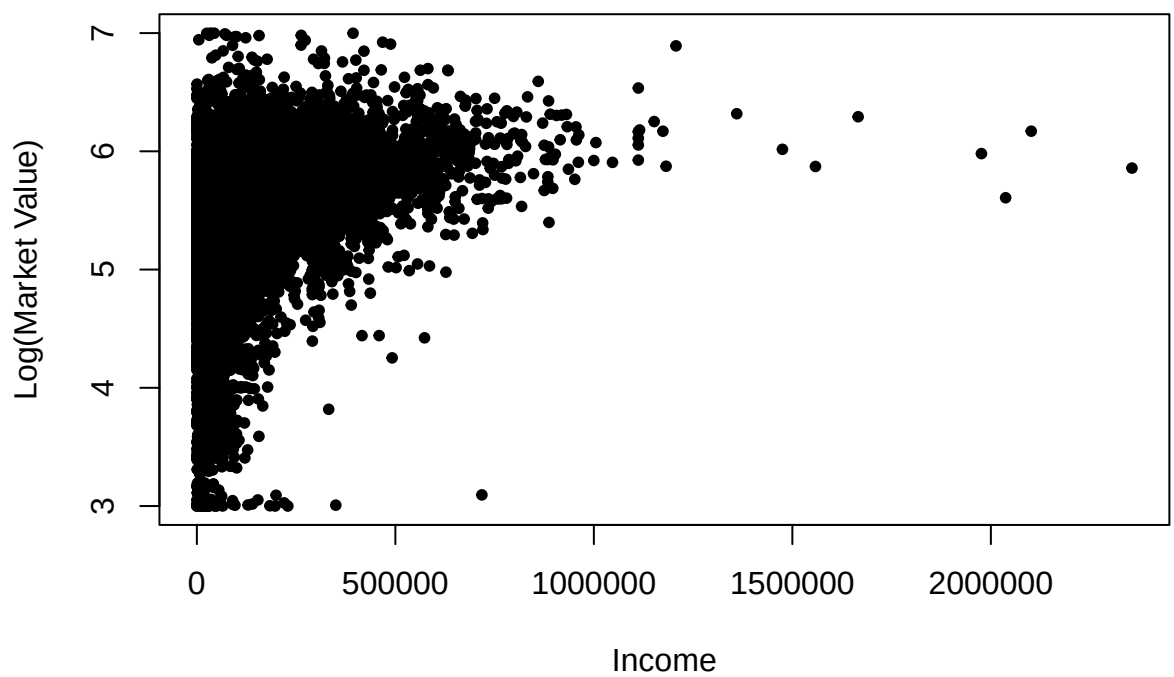


Since both variables are highly skewed to the right, it is hard to discern a pattern, so we'll try and re-express the data by taking logarithms.

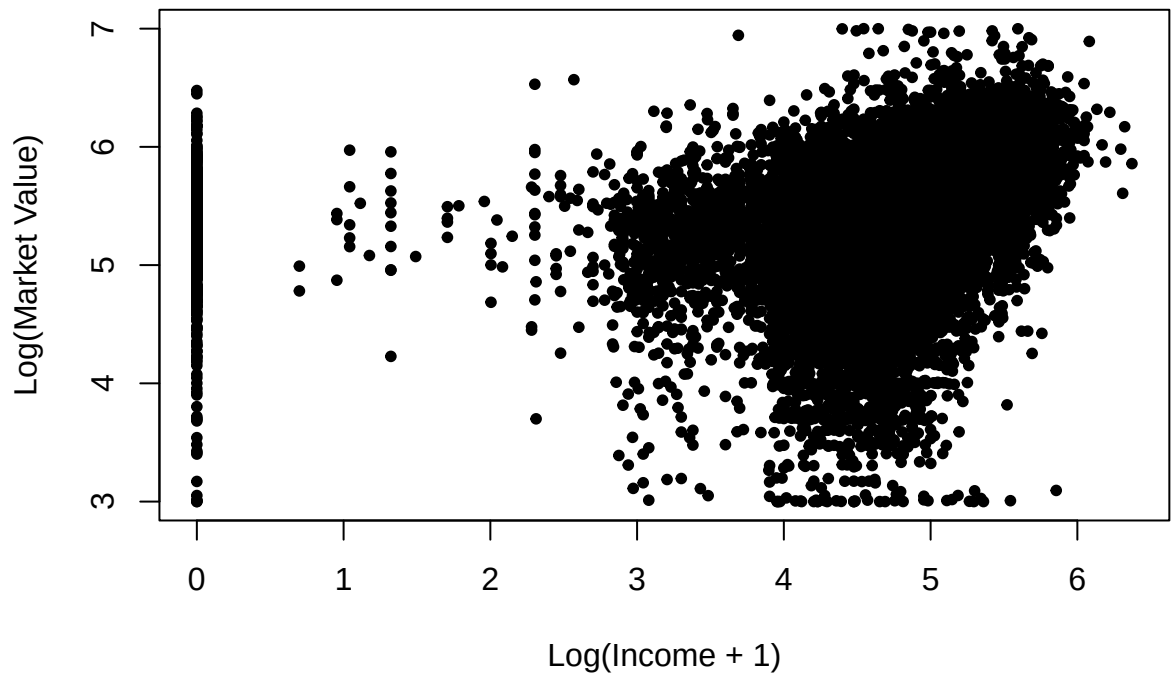
```
log_income = log(income+1, 10)
plot(log_income, market_val,
     xlab = "Log(Income + 1)", ylab="Market Value", pch = 20)
```




```
log_market_val = log(market_val, 10)
plot(income, log_market_val,
     xlab = "Income", ylab="Log(Market Value)", pch = 20)
```

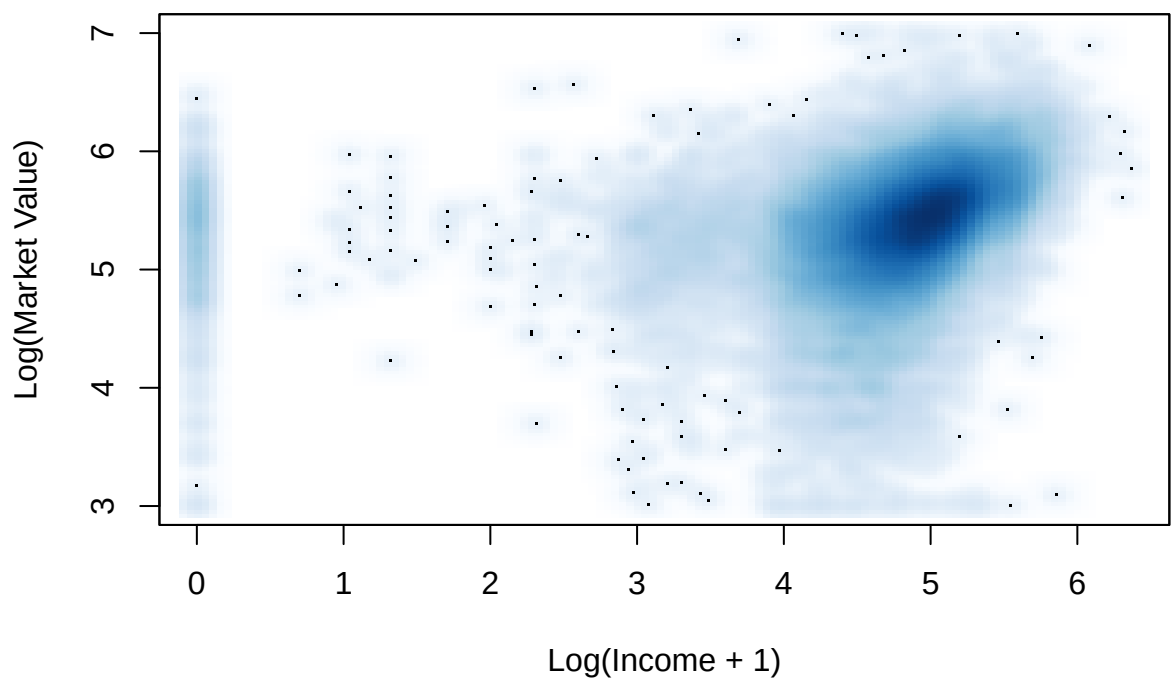



```
log_income = log(income+1, 10)
plot(log_income, log_market_val,
     xlab = "Log(Income + 1)", ylab = "Log(Market Value)", pch=20)
```



Here's a "heat map" style version of the scatterplot, to show where the points are concentrated:

```
smoothScatter(log_income, log_market_val, xlab = "Log(Income + 1)", ylab = "Log(Market
↪  Value)")
```



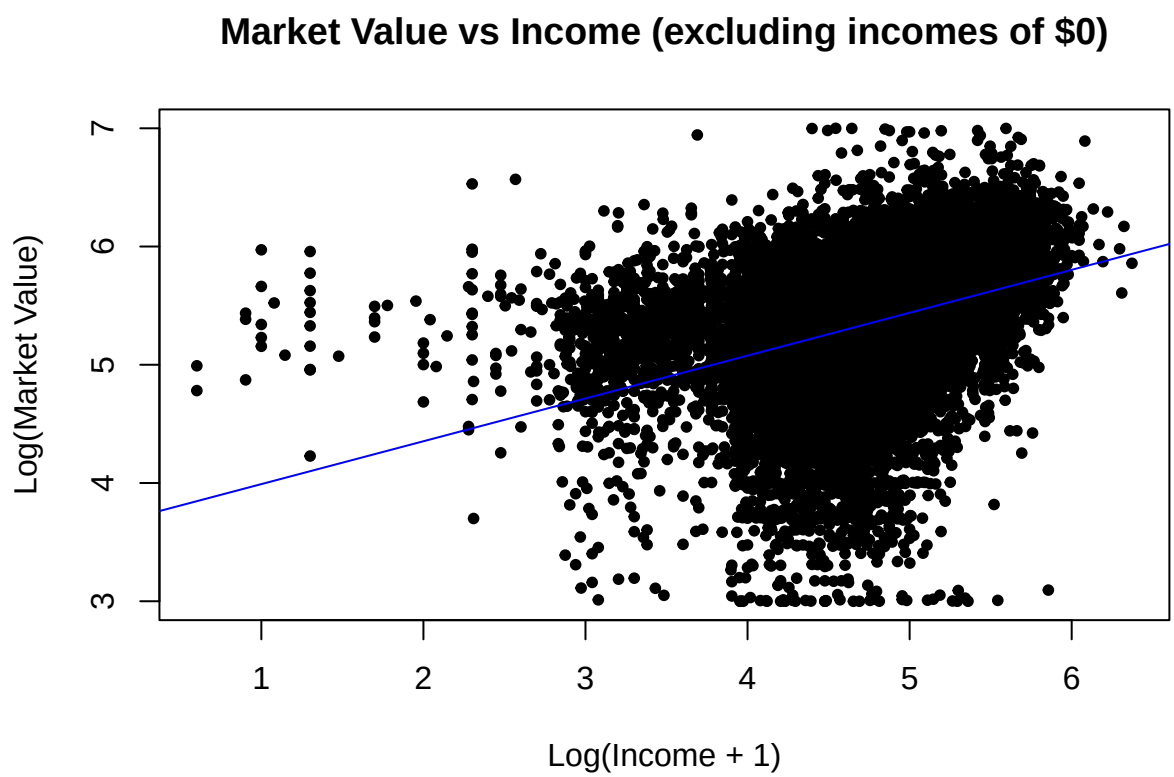
Clearly, most points are concentrated at Log(Income + 1) values near 5 (which is represents about $100,000 in annual income), and Log(Market Value) values between 5 and 6, which corresponds to house prices in the range of $100,000 to $1,000,000.

Notice the veritcal line on the left, which represents all the households with reported annual incomes of $0. Despite apparently having no income, some of these households have houses whose market value is over $1,000,000! This may indicate that many of these households are exploiting loopholes when filing taxes, falsely reporting their incomes as zero. Of course, this is just a hypothesis, and cannot be directly concluded

from this data. Further investigation into the finances of these households would be required to confirm or deny this hypothesis.

Since this vertical bar would make the slope of a regression line lower than it would be otherwise, we'll remove these houses when performing our analysis. Although this scatterplot does not seem to satisfy the "Straight Enough Condition", we'll proceed with linear regression analysis anyway.

```
data_positive_income <- data[!(data$FINCP <= 0),]
log_income_no_zero <- log(data_positive_income$FINCP, 10)
log_marketval_income_no_zero <- log(data_positive_income$MARKETVAL, 10)


reg_marketval_vs_income <- lm(log_marketval_income_no_zero ~ log_income_no_zero)
plot(log_marketval_income_no_zero ~ log_income_no_zero,
     main = "Market Value vs Income (excluding incomes of $0)", xlab = "Log(Income + 1)",
     ↪  ylab = "Log(Market Value)", pch=20)
abline(reg_marketval_vs_income, col="blue")
```



The summary statistics associated with this regression are shown below.

```
summary(reg_marketval_vs_income)
```

```
Call:
lm(formula = log_marketval_income_no_zero ~ log_income_no_zero)

Residuals:
     Min       1Q   Median       3Q      Max
-2.65669 -0.18491  0.02677  0.22799  2.06743

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
```

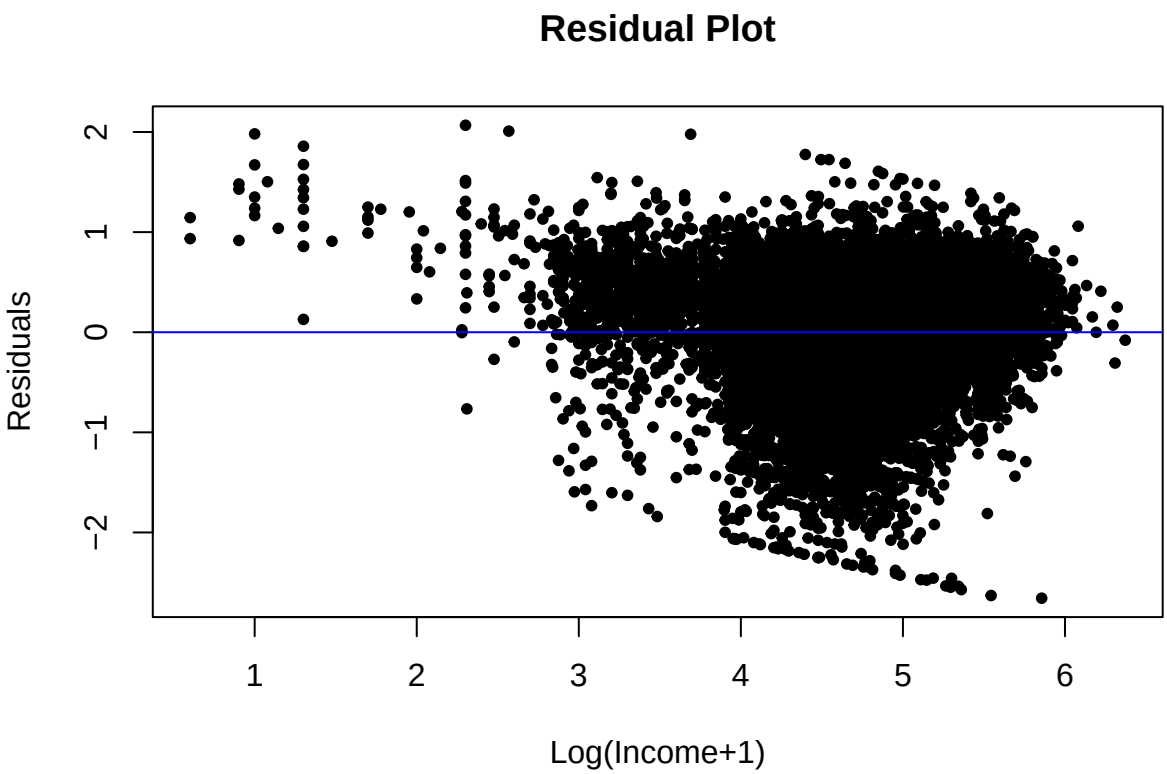
```
(Intercept)        3.628058   0.023179   156.5   <2e-16 ***
log_income_no_zero 0.362485   0.004776    75.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4083 on 32472 degrees of freedom
Multiple R-squared:  0.1507,    Adjusted R-squared:  0.1507
F-statistic:  5761 on 1 and 32472 DF,  p-value: < 2.2e-16
```

The equation of the least squares regression line is

$$\widehat{\log(\text{MarketValue})} = 3.628 + 0.362 \cdot \log(\text{Income} + 1)$$

The residual plot is shown below:

```r
# Residual Plot
plot(log_income_no_zero, resid(reg_marketval_vs_income),
     main="Residual Plot", xlab="Log(Income+1)", ylab="Residuals", pch=20)
abline(0,0, col="blue")
```



The residual plot is not ideal: It still looks like there is an uneven spread about the horizontal line, and a slight downward trend. However, no re-expression we have tried has resulted in a better residual plot.

Now, let's study the statistics relevant to the regression analysis.

```
The slope of the least-squares regression line is 0.362.
```

```
This implies that each additional unit increase in Log(Income) is associated with an increase of 0.362 u
```

In other words, each additional increase in income by a factor of 10 is associated with an increase in Market Value by a factor of 10^{0.362}, i.e., an increase in market value by a factor of about 2.30.

```
The correlation coefficient, r, is 0.388
```

Since r is positive, there is a positive association between Log(Market Value) and Log(Income+1). However, since r is relatively low, there is a *weak* linear correlation between Log(Market Value) and Log(Income+1), which makes sense given how messy the scatterplot is. This specific r value means that, for every one standard deviation above (or below) the mean a household's Log(Income+1) is, we predict that its Log(Market Value) is 0.388 standard deviations above (or below) its mean Log(Market Value).

Now, let's look at the R-squared value.

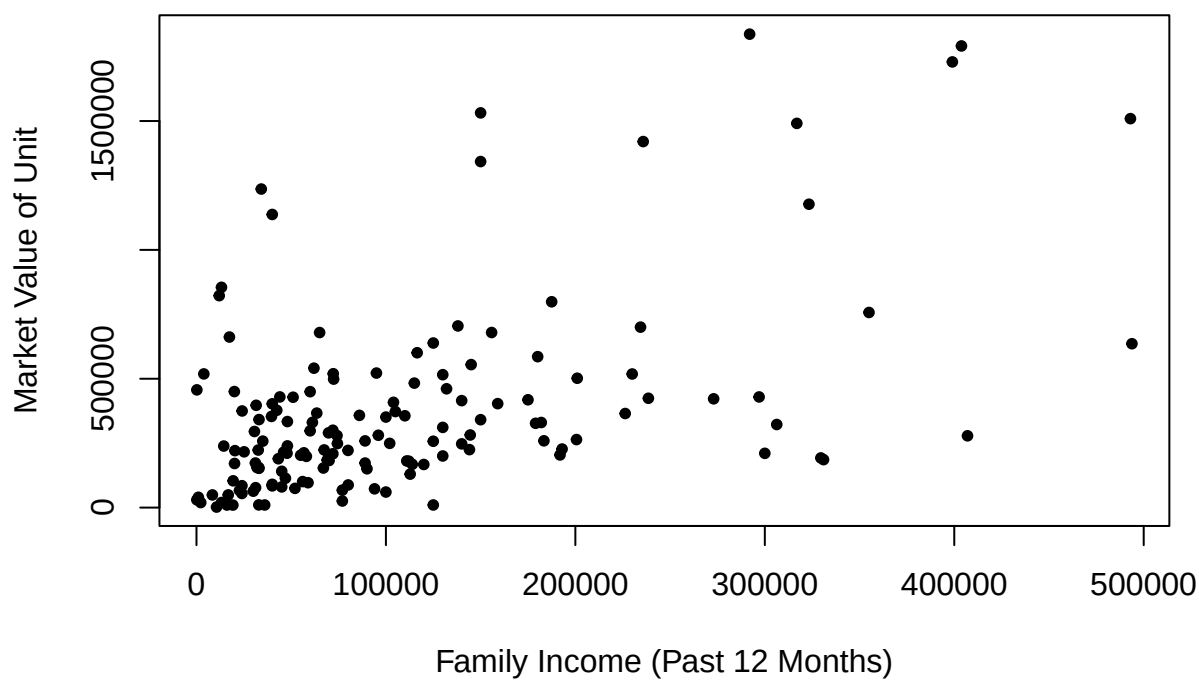
```
The R-squared value is 0.151
```

This means that about 15.1% of the variability in Log(Market Value) is explained by the model.

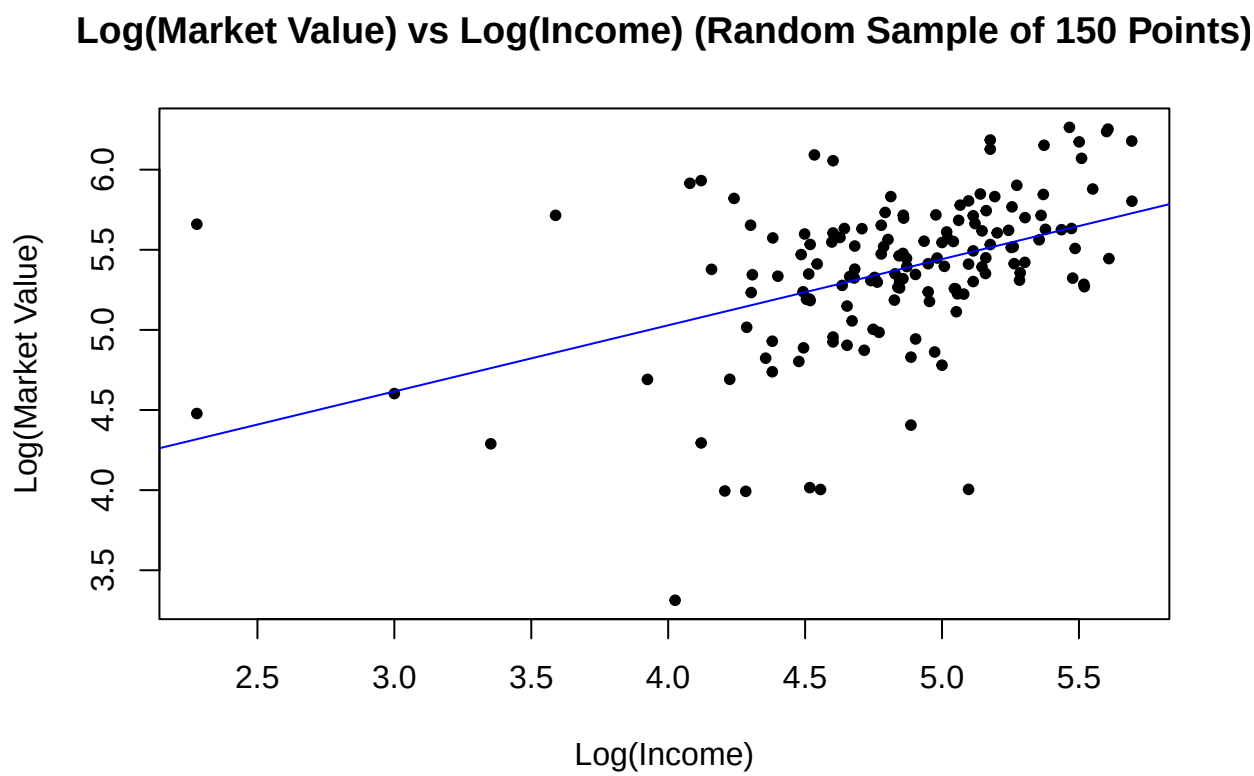
Considering a smaller sample

Since it is hard to discern a pattern from this large dataset, we'll perform regression analysis again on a random sample of 100 elements from the dataset to better demonstrate learning.

```r
# Simple Random Sampling
set.seed(100) # This can be any number
random_subset_of_data <- data_positive_income[sample(nrow(data_positive_income), 150),]
random_marketval <- random_subset_of_data$MARKETVAL
random_income <- random_subset_of_data$FINCP
plot(random_marketval ~ random_income,
     xlab = "Family Income (Past 12 Months)", ylab = "Market Value of Unit", pch=20)
```




```r
log_random_marketval <- log(random_marketval, 10)
log_random_income <- log(random_income, 10)
reg_subset <- lm(log_random_marketval ~ log_random_income)
plot(log_random_marketval ~ log_random_income,
     main = "Log(Market Value) vs Log(Income) (Random Sample of 150 Points)", xlab =
     ↪  "Log(Income)", ylab = "Log(Market Value)", pch=20)
abline(reg_subset, col="blue")
```



The summary statistics associated with this regression are shown below.

```
summary(reg_subset)
```

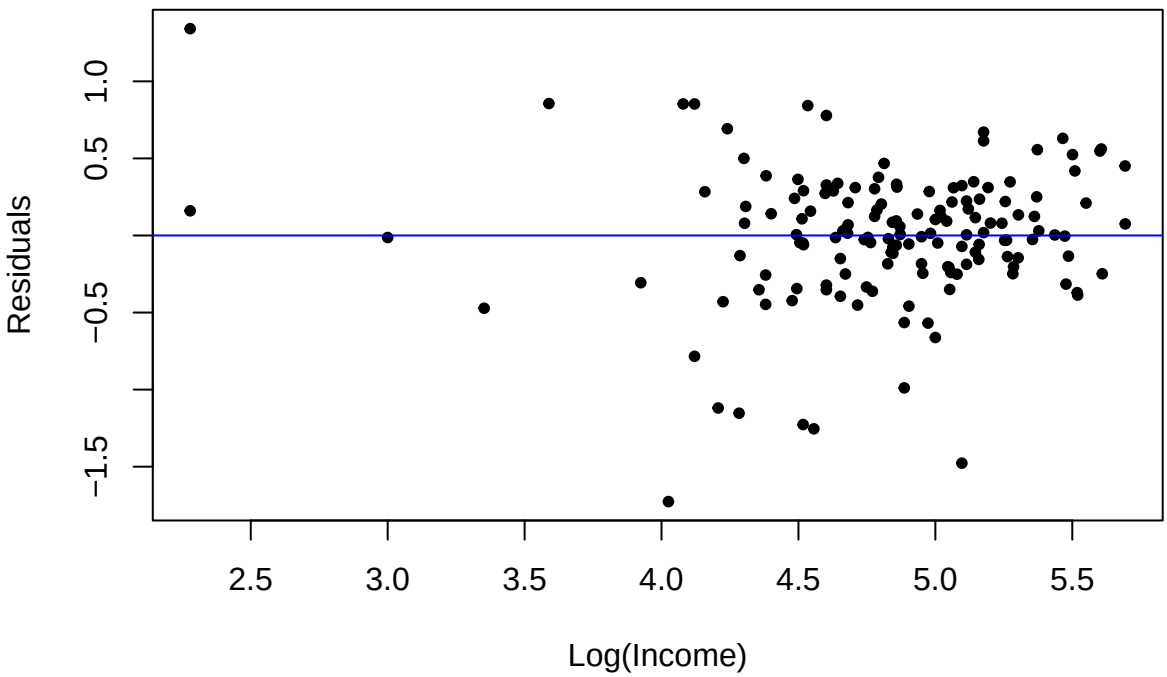
```
Call:
lm(formula = log_random_marketval ~ log_random_income)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.72641 -0.20318  0.01016  0.24829  1.34172 

Coefficients:
                  Estimate Std. Error t value     Pr(>|t|)    
(Intercept)        3.37696    0.32763  10.307      < 2e-16 ***
log_random_income  0.41295    0.06756   6.112 0.0000000083 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4418 on 148 degrees of freedom
Multiple R-squared:  0.2016,	Adjusted R-squared:  0.1962 
F-statistic: 37.36 on 1 and 148 DF,  p-value: 0.000000008304
```

The equation of the least squares regression line is

$$\log(\widehat{\text{MarketValue}}) = 3.377 + 0.413 \cdot \log(\text{Income})$$

The residual plot is shown below:

```
# Residual Plot
plot(log_random_income, resid(reg_subset),
     xlab="Log(Income)", ylab="Residuals", pch=20)
abline(0,0, col="blue")
```



This is a fairly good residual plot. Although the points are concentrated to the right, the spread along the horizontal is fairly even.

Now, let's study the statistics relevant to the regression analysis.

```
The slope of the least-squares regression line is 0.413.
```

```
This implies that each additional unit increase in Log(Income) is associated with an increase of 0.413 u
```

In other words, each additional increase in income by a factor of 10 is associated with an increase in Market Value by a factor of 10^{0.413}, i.e., an increase in market value by a factor of about 2.588.

```
The correlation coefficient, r, is 0.449
```

Since r is positive, there is a positive association between Log(Market Value) and Log(Income). The r here is a bit larger than the r value for the original scatterplot with all 30,000 points (recall: the original dataset had a correlation coefficient of 0.388). Thus, there is a stronger linear correlation shown for this specific random sample of the dataset than there was for the overall dataset. This specific r value means that, for every one standard deviation above (or below) the mean a household's Log(Income) is, we predict that its Log(Market Value) is 0.449 standard deviations above (or below) its mean Log(Market Value).

Now, let's look at the R-squared value.

```
The R-squared value is 0.202
```

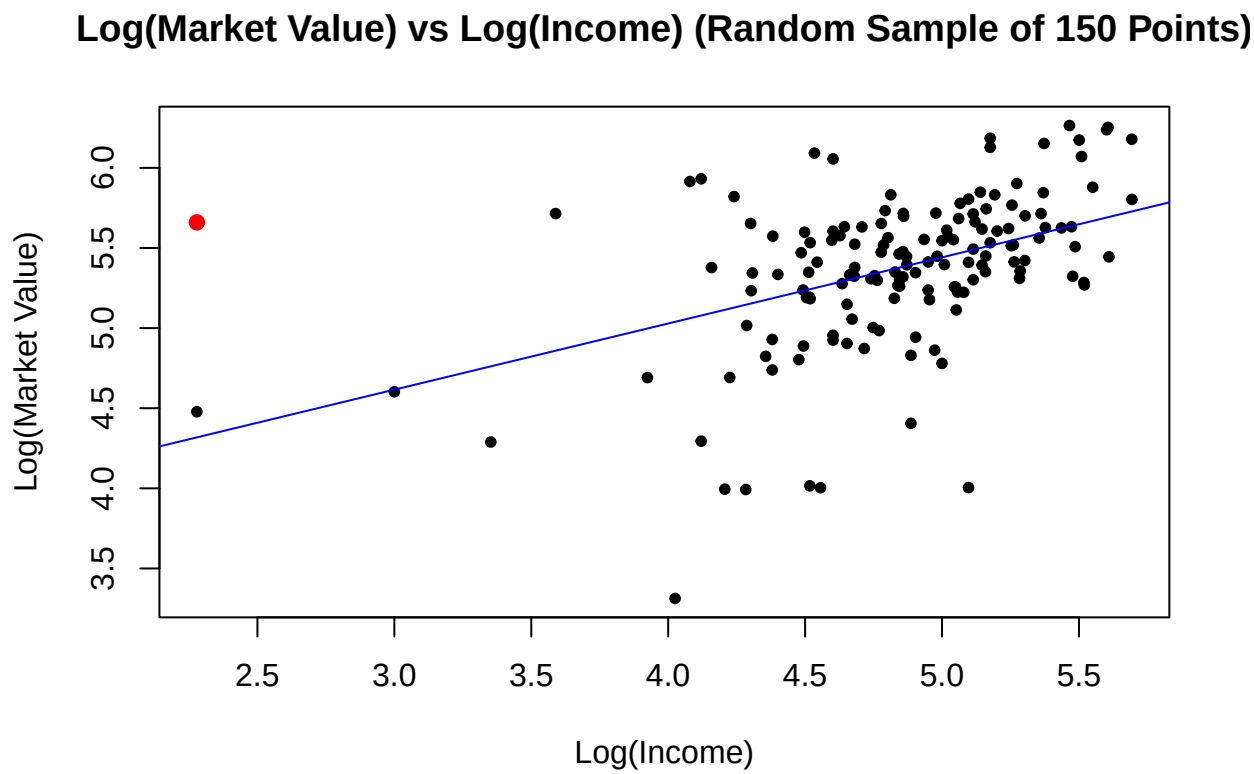
This means that about 20.2% of the variability in Log(Market Value) is explained by the model. This is a bit better than the R^2 value for the full dataset, which is reasonable since the variability shown in the original dataset is much greater than the variability shown in the scatterplot of this random subset of the data.

Investigating the Effects of Outliers

Once again, here is the scatterplot of this random subset of data:

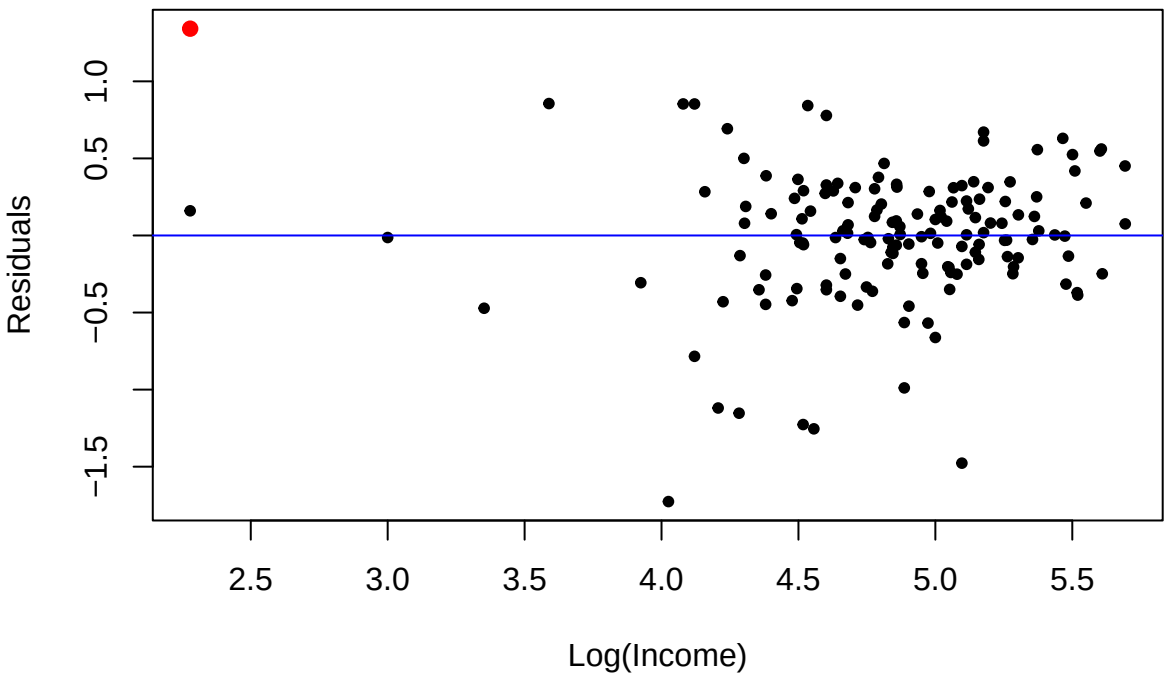
```
colors <- "black"
point_style <- 20
for (i in 2:150) {
  colors <- c(colors, "black")
  point_style <- c(point_style, 20)
```

```
}
colors[36] <- "red"
point_style[36] <- 19
plot(log_random_marketval ~ log_random_income,
     main = "Log(Market Value) vs Log(Income) (Random Sample of 150 Points)", xlab =
     ↪  "Log(Income)", ylab = "Log(Market Value)",
     col = colors, pch=point_style)
abline(reg_subset, col="blue")
```



Notice the point labelled with the color red in the top left. This is an outlier with high leverage since its Log(Income) value appears to be far to the left of the mean Log(Income), which is easy to see since the rest of the data is clumped towards the right side of the plot. It also has a high residual, which can be seen by marking it on the residual plot.

```
# Residual Plot
plot(log_random_income, resid(reg_subset),
     xlab="Log(Income)", ylab="Residuals",
     col = colors, pch=point_style)
abline(0,0, col="blue")
```

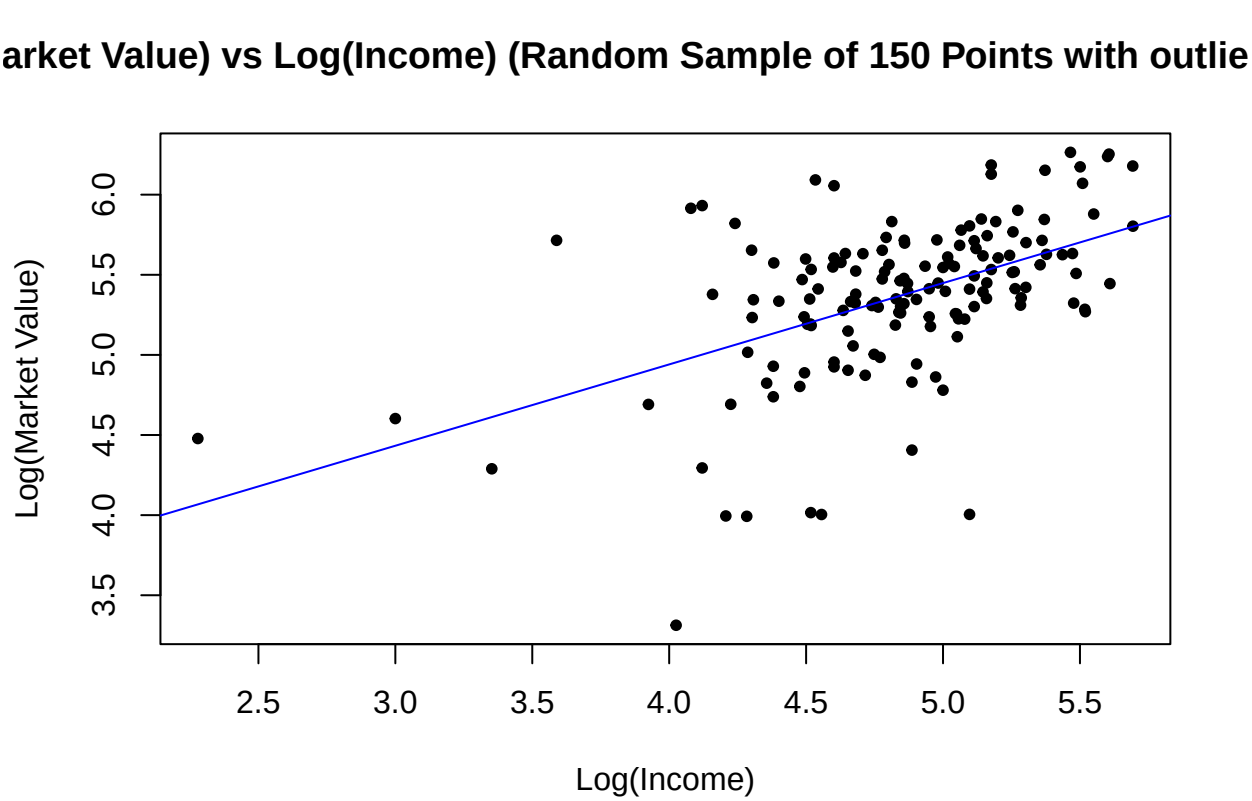



```
message("The residual of the red outlier is about ", round(resid(reg_subset)[36], 3))
```

```
The residual of the red outlier is about 1.342
```

To check if this point is influential, regression analysis must be performed yet again — this time with the outlier ommitted.

```
log_random_marketval_no_outlier <- c(log_random_marketval[1:35],
↪  log_random_marketval[37:150])
log_random_income_no_outlier <- c(log_random_income[1:35], log_random_income[37:150])
reg_subset_no_outlier <- lm(log_random_marketval_no_outlier ~
↪  log_random_income_no_outlier)
plot(log_random_marketval_no_outlier ~ log_random_income_no_outlier,
     main = "Log(Market Value) vs Log(Income) (Random Sample of 150 Points with outlier
     ↪  removed)", xlab = "Log(Income)", ylab = "Log(Market Value)", pch=20)
abline(reg_subset_no_outlier, col="blue")
```



The summary statistics associated with this regression are shown below.

```
summary(reg_subset_no_outlier)
```

```
Call:
lm(formula = log_random_marketval_no_outlier ~ log_random_income_no_outlier)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.64059 -0.21118  0.00915  0.28082  0.98304 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  2.91021    0.34443   8.449 2.62e-14 ***
log_random_income_no_outlier  0.50759    0.07084   7.165 3.48e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4266 on 147 degrees of freedom
Multiple R-squared:  0.2588,	Adjusted R-squared:  0.2538 
F-statistic: 51.34 on 1 and 147 DF,  p-value: 3.477e-11
```

The equation of the least squares regression line is

$$\log(\widehat{\text{MarketValue}}) = 2.910 + 0.508 \cdot \log(\text{Income})$$

The previous equation had a y-intercept of 3.377, while this one has a y-intercept of 2.910. Clearly, the outlier had been pulling the y-intercept higher. Also, the previous equation had a slope of 0.413, while this one has a slope of 0.508. Thus, the upward pull of this outlier had been making the line more flat.

```
The correlation coefficient, r, is 0.509
```

This correlation coefficient is greater than the previous correlation coefficient (which was 0.449). Thus, omitting this outlier results in a stronger linear correlation between market value and income.

```
The R-squared value is 0.259
```

The R^2 value for this regression line is greater than the R^2 value with the outlier included (which was 0.202). With the outlier removed, 25.9% of the variability in market value is explained by the model.

Clearly, this outlier was influential, since because removing it from the regression analysis resulted in a noticeably different model. This is shown by the visual difference in the trajectory of the least-squares regression line, the differences in the sloper and y-intercept of the regression equation, and the notable increase in the correlation coefficient and R^2 value.

Conclusion

Overall, our original hypothesis that there is a positive correlation between the market value of a family's house and the family's annual income is supported to some extent. There does indeed seem to be a positive association between market value and income, as shown by the positive slope, correlation coefficient, and shape of the graph.

However, we had been expecting a moderate or strong linear correlation after re-expression, but no form of re-expression generated a graph that was irrefutably "Straight Enough" and didn't violate the "Does the Plot Thicken?" condition, since the data was concentrated around a specific region (as shown by the dark spot in the "heat map"). These issues were solved to some extent by selecting a random sample of 150 points from the data and repeating the analysis. However, this method is not ideal since it attempts to better understand the data by *decreasing* the sample size, rather than *increasing* it.

One possible next step would be to try to gather more data for households with more extreme income values, which could be done by adding more rich neighborhoods and poor neighborhoods to the survey pool. This would even out the spread along the income spectrum, reducing the high concentration of data around the $100,000 - $1,000,000 income range. This would in turn make the spread along the least-squares regression line more even, which might make the data seem more straight and satisfy the "Does the Plot Thicken?" condition.

The low R^2 value of 0.151 in the original regression (over 30,000 data points) indicated that still 85% of the variation in Log(Market Value) was left unexplained by the model. A model that takes into account more factors (like household demographics, location, unit size, etc.) would likely do a better job predicting market value.

Source

United States Census Bureau. American Housing Survey 2019 National Public Use File. Retrieved from https://www.census.gov/programs-surveys/ahs/data/2019/ahs-2019-public-use-file--puf-/ahs-2019-national-public-use-file--puf-.html.