# California State University Long Beach

Department of Mathematics
Applied Statistics

STAT Regression Analysis
Spring 2023

*Regression Analysis of Housing Prices*

Luis Osorio, Richard Diazdeleon, Victor Chen

# Regression Analysis of Housing Prices

Luis Osorio, Richard Diazdeleon, Victor Chen

September 5, 2023

## 1 Introduction

The USA Housing data-set[1], available on Kaggle, contains information on various housing parameters for the different cities in the United States of America. The data set includes features like average house age, average rooms, average bedrooms, population, average income, average house price, and more. The data-set consists of 5000 rows and 7 columns, and it is in CSV (Comma Separated Values) format. Housing prices in the United States have increased over time and we want to see what kind of variables are most important when looking at purchasing a house as an investment. The surrounding neighborhood and city can affect house prices and the data set contains many of those features. The regression model has value when submitting a bid on a house for future homeowners.

Some common attributes about the data are as followed:

- 5000 rows
- ‘Avg. Area Income’: Average Income of residents of the city house is located in.
- ‘Avg. Area House Age’: Average Age of Houses in same city
- ‘Avg. Area Number of Rooms’: Average Number of Rooms for Houses in same city
- ‘Avg. Area Number of Bedrooms’: Average Number of Bedrooms for Houses in same city
- ‘Area Population’: Population of city house is located in
- ‘Price’: Price that the house sold at
- ’Address’: Address for the house

## 2 Question of Interest

- What combination of features (income, age of houses, number of rooms, and number of bedrooms) are statistically significant in predicting the selling price of a house?
- Is the reduced or full linear regression model best suited to predict housing prices?
- What is the average house price in the united states when the average area income is $50,000, average area house age is 5, an area population of 35,000 people, and average area number of rooms is 6?

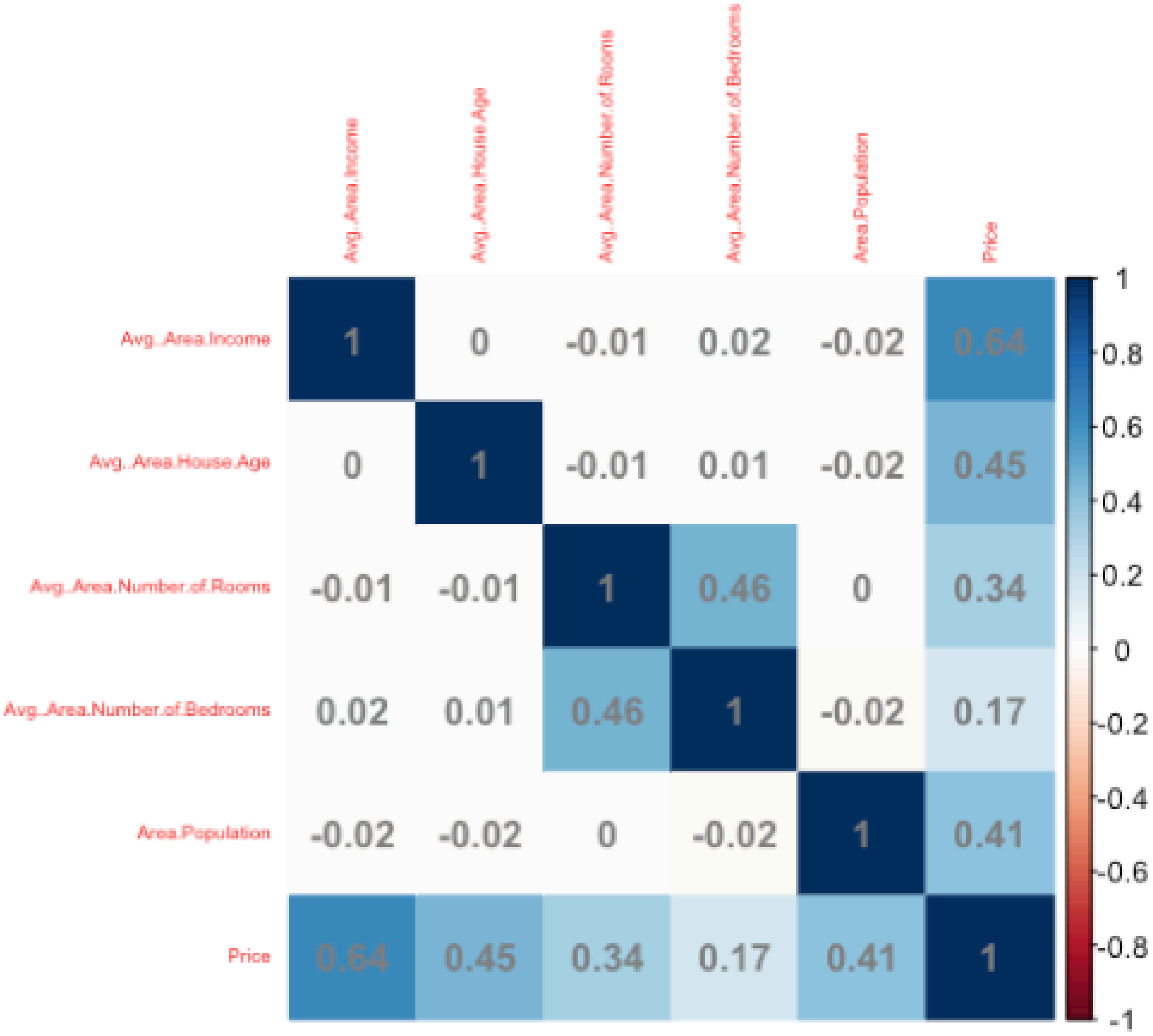
## 3 Regression Method

To answer the first research question in determining the most significant predictors we will use the t-test on the slope parameters. We also plan to explore the results further using the overall F-test to check if we get similar results. Next we check the significance of the interactions between our numeric variables and see if our initial model improves. We will also use the general linear F-test to compare the two models. After checking all assumptions for our residuals and finding the best model, we will use this model to make a type of prediction called confidence.

# 4 Regression Analysis, Results and Interpretation

## 4.1 Correlations on our Response Variable

We introduce the following data by a given scatter-plot matrix to try to identify which predictors are highly correlated to price. As a result, we have the following heat-map on our correlation matrix

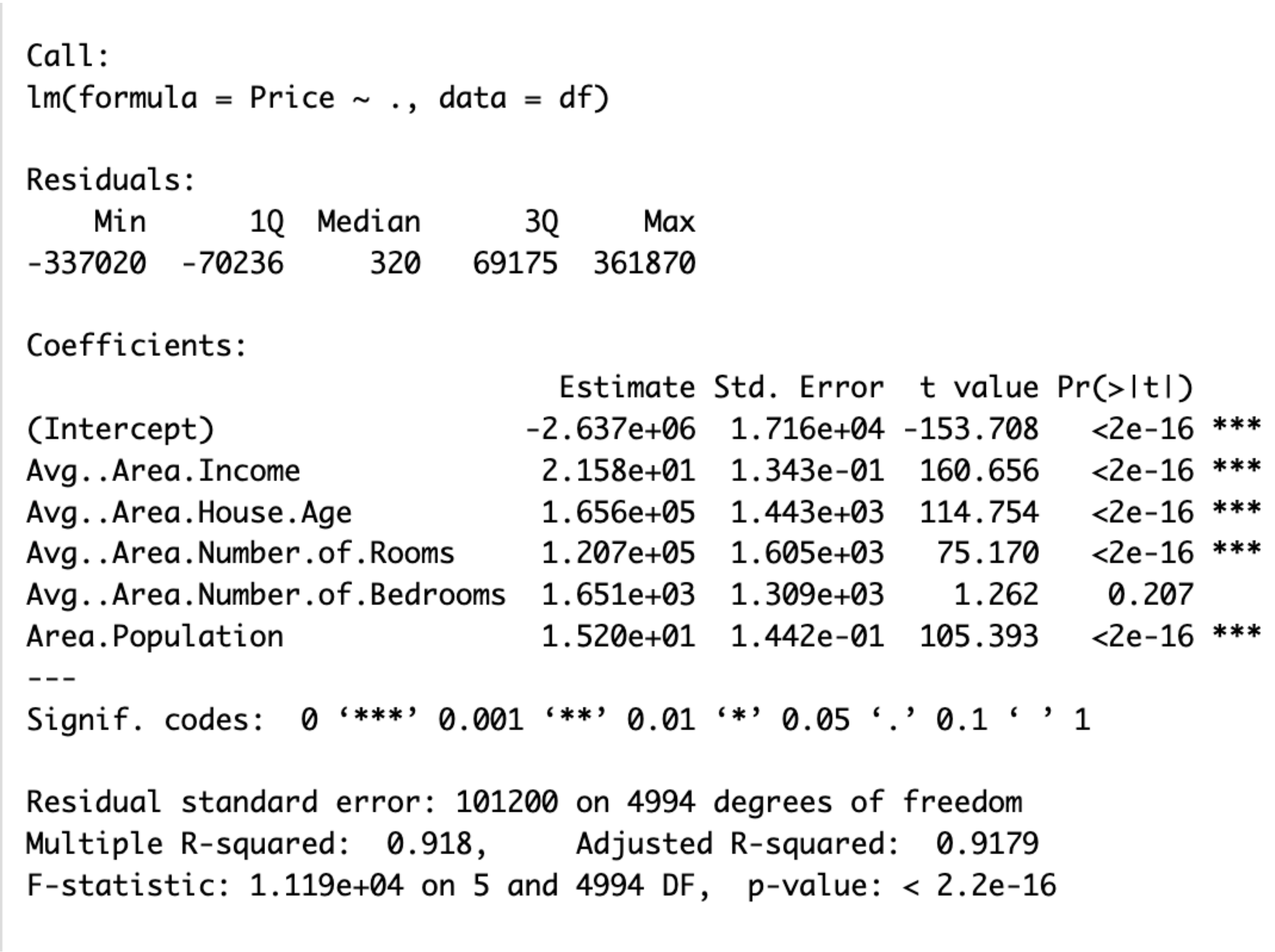


| | Avg..Area.Income | Avg..Area.House.Age | Avg..Area.Number.of.Rooms | Avg..Area.Number.of.Bedrooms | Area.Population | Price |
|---|---|---|---|---|---|---|
| Avg..Area.Income | 1 | 0 | -0.01 | 0.02 | -0.02 | 0.64 |
| Avg..Area.House.Age | 0 | 1 | -0.01 | 0.01 | -0.02 | 0.45 |
| Avg..Area.Number.of.Rooms | -0.01 | -0.01 | 1 | 0.46 | 0 | 0.34 |
| Avg..Area.Number.of.Bedrooms | 0.02 | 0.01 | 0.46 | 1 | -0.02 | 0.17 |
| Area.Population | -0.02 | -0.02 | 0 | -0.02 | 1 | 0.41 |
| Price | 0.64 | 0.45 | 0.34 | 0.17 | 0.41 | 1 |

and so by looking at our Price column we can immediate note that the ’Avg.Area.Income’ is a strong predictor with a score of 0.64 and in result the correlation matrix shows which predictors may be significant. Investigating the matrix further we also observed that ’Average Area Number of Rooms’ and ’Average Area Number of Bedrooms’ are also highly correlated with a score of 0.46. When two predictors are highly correlated it signifies that one variable contains similar information to the other variable. This is a for shadow of what may be the significant predictors.

## 4.2 Residual Analysis

Using the lm() function in R and we fit a regression model using all predictors. Then we analyze our summary table to get further insights about the fitted model.

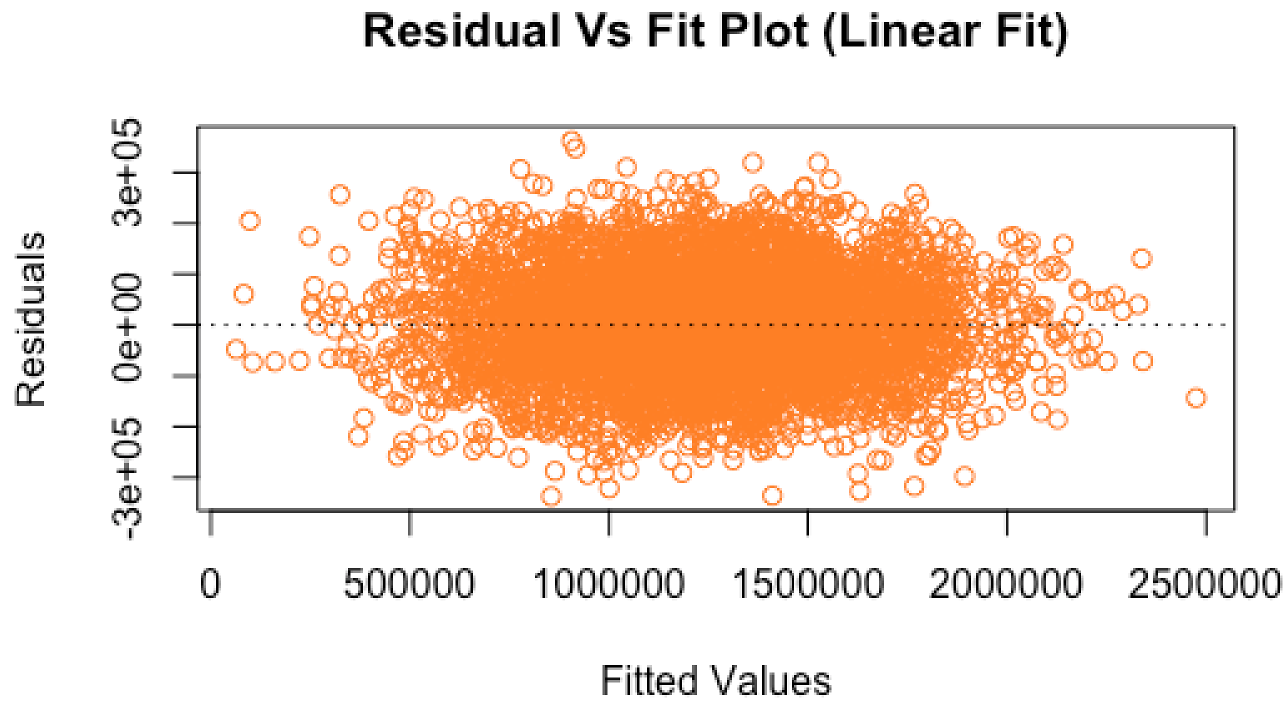


```
Call:
lm(formula = Price ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-337020  -70236     320   69175  361870 

Coefficients:
                               Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                  -2.637e+06  1.716e+04 -153.708   <2e-16 ***
Avg..Area.Income              2.158e+01  1.343e-01  160.656   <2e-16 ***
Avg..Area.House.Age           1.656e+05  1.443e+03  114.754   <2e-16 ***
Avg..Area.Number.of.Rooms     1.207e+05  1.605e+03   75.170   <2e-16 ***
Avg..Area.Number.of.Bedrooms  1.651e+03  1.309e+03    1.262    0.207    
Area.Population               1.520e+01  1.442e-01  105.393   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101200 on 4994 degrees of freedom
Multiple R-squared:  0.918,	Adjusted R-squared:  0.9179 
F-statistic: 1.119e+04 on 5 and 4994 DF,  p-value: < 2.2e-16
```

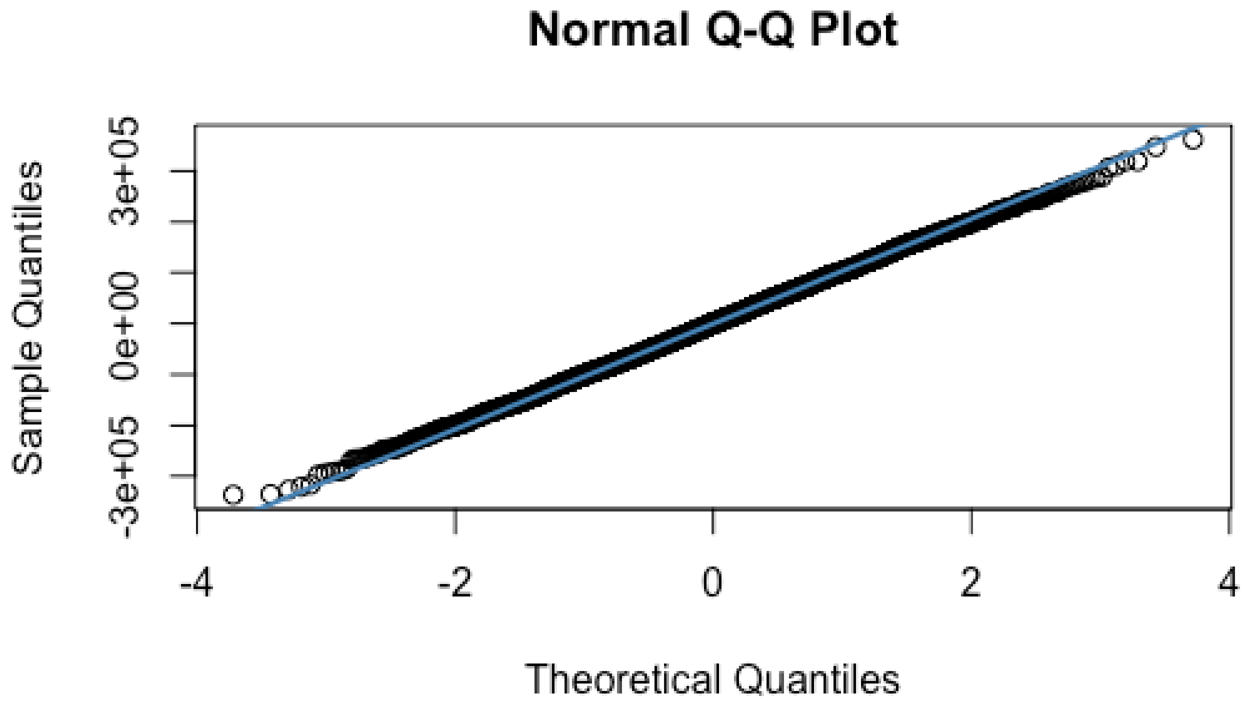
Our fitted linear model has Price as the response variable and the other variables are the predictors. Using the summary table we can get the predictors coefficients estimates to get our first predicted linear model. Where our population model is written as,

$$Price = \beta_0 + \beta_1 Avg..Area.Income + \beta_2 Avg..House.Ag + \beta_3 Avg..Number.of.Rooms + \beta_4 Area.Population + \epsilon_i$$

Now we want to check if this model was plausible by checking the four LINE conditions and so we begin by checking the residuals plot. The Four conditions that we must check are independence, linearity, normality, equal variance on the errors. The residuals plot consist of having the residuals plotted along the y-axis and the fitted values along the x-axis.



The residual analysis was performed on the linear model with all variables as predictors. The residual versus fit plot was examined, which suggested that the equal variance assumption across the residuals was met since the points fit inside a horizontal band. Additionally, it was observed that the linearity assumption was met, as the points appeared to bounce randomly around the zero line. Next we check for the the normality condition on the errors. This can be done using several methods such as QQ-plot, Shapiro-Test, or using a histogram. Using a QQ-plot, we need to verify if our points follow consistently upon our theoretical line. Thus, we can analyze our



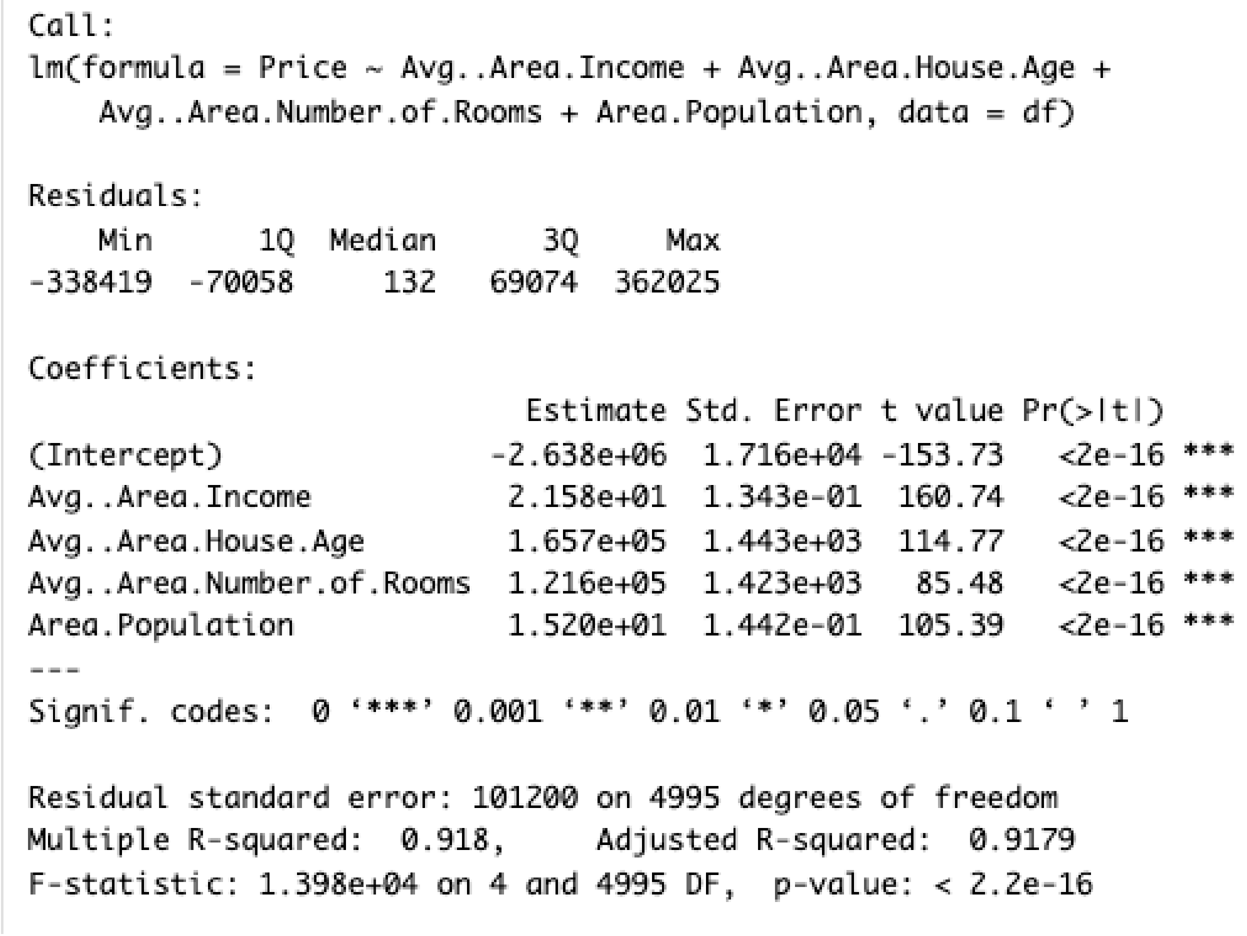
normality assumption on the residuals and note that the normality assumption was satisfied using a QQ-plot. This indicates that the errors are normally distributed, and therefore the linear regression model with all variables as predictors is appropriate for answering the research questions.

## 4.3 Q1. What combination of features (income, age of houses, number of rooms, and number of bedrooms) are statistically significant in predicting the selling price of a house?

We want to first fit each of the predictors in our model where Y regresses on all x. Next, for each of those predictors we want to be able to test if the predictor has a relationship with Y, thus setting up our hypothesis test. Recall, that the t-test checks the relationship of each predictor one by one. Thus, our null hypothesis will imply that the predictor has no relationship with the response and our alternative will imply there is a relationship. Our hypothesis test can be summarized mathematically down below.

$$H_0 : B_1 = 0 \text{ vs } H_1 : B_1 \neq 0$$

Now that we have our test, we now have to set a threshold to check if our assumptions meets it or falls below it. Our significance level, $\alpha = 0.05$ where if our p-value falls below this threshold we will reject our null hypothesis and conclude that the predictor is significant to the model. However, if the p-value is greater than our threshold than we will fail to reject the null hypothesis and state that the predictor has no relationship with our response variable. No relationship to the response will suggest a insignificant predictor therefore we can remove the predictor to produce an efficient model. After running the results we observed that that only variable that failed to fall below our threshold was ’Average Number of Bedrooms’ implying it is statistically insignificant to our model. So after determining the significant predictors we fitted a new linear regression model with only the significant predictors. The new model summary table is down below,



```
Call:
lm(formula = Price ~ Avg..Area.Income + Avg..Area.House.Age +
    Avg..Area.Number.of.Rooms + Area.Population, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-338419  -70058     132   69074  362025

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.638e+06  1.716e+04 -153.73   <2e-16 ***
Avg..Area.Income          2.158e+01  1.343e-01  160.74   <2e-16 ***
Avg..Area.House.Age       1.657e+05  1.443e+03  114.77   <2e-16 ***
Avg..Area.Number.of.Rooms 1.216e+05  1.423e+03   85.48   <2e-16 ***
Area.Population           1.520e+01  1.442e-01  105.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101200 on 4995 degrees of freedom
Multiple R-squared:  0.918,     Adjusted R-squared:  0.9179
F-statistic: 1.398e+04 on 4 and 4995 DF,  p-value: < 2.2e-16
```

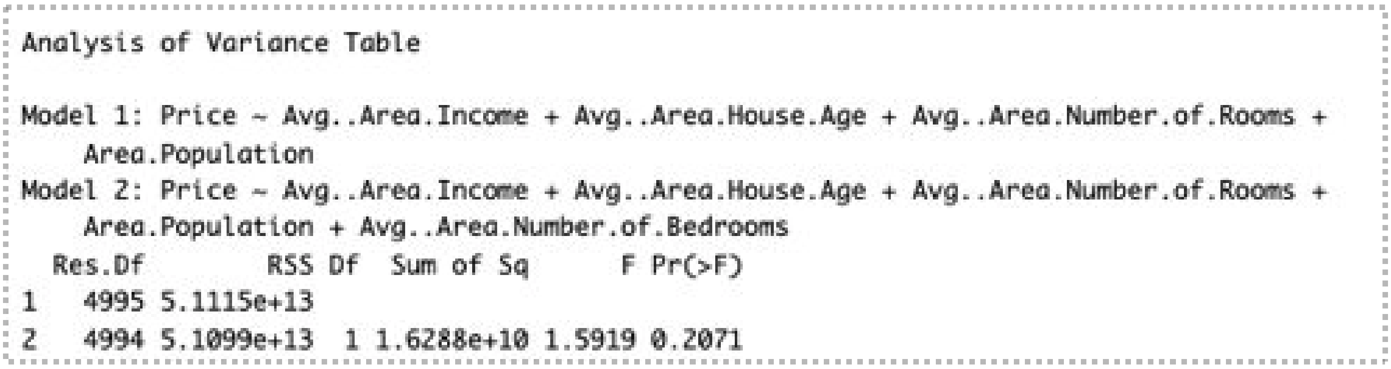
Now comparing the summary tables between the two models where the first model used all predictors and the second model used only the principle variables, we hardly noticed any differences. For example, our $R^2$ value remained the same between the model which proved further that dropping the variable did not improve our model at all. Our new expected linear regression model can be written as

$$Price = -2.638e^{06} + 2.158e^{1}Avg..Area.Income$$

$$+1.657e^{5}Avg..House.Age + 1.216e^{5}Avg..Number.of.Rooms + 1.520e^{1}Area.Population.$$

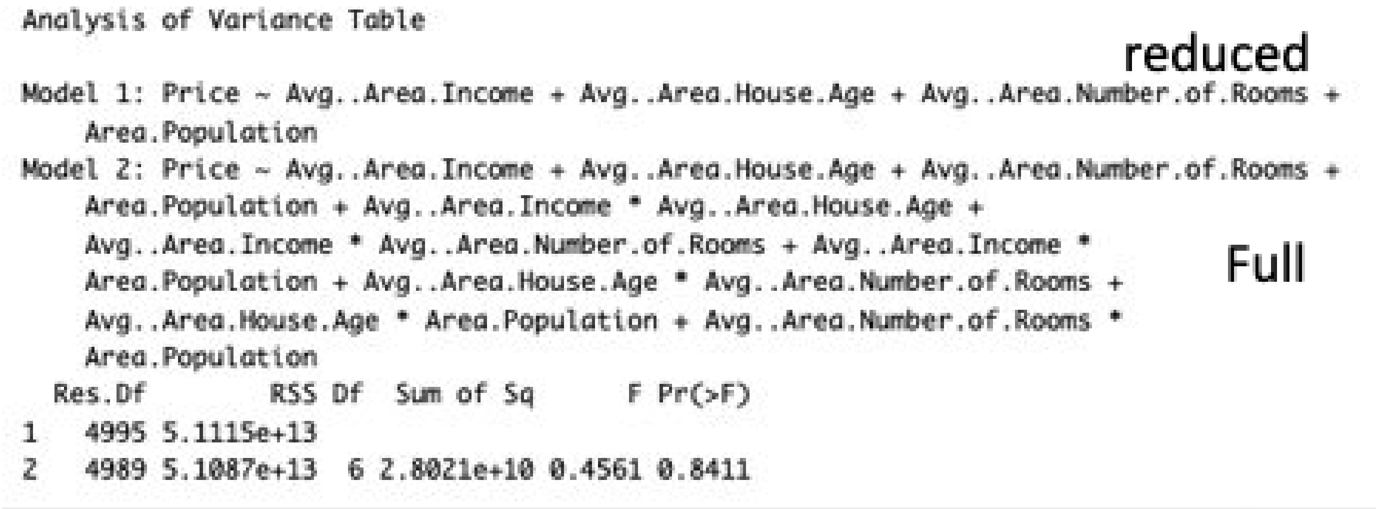
## 4.4 Q2. Is the reduced or full linear regression model best suited to predict housing prices?

The goal here will be to use a general linear F-test to check for interactions significance between the important features. After conducting a T-test on all the variables, we were able to conclude that the only insignificant predictor was the 'Average Area of Number of Bedroom'. It was removed from our model but now we will check this by doing the overall f-test to check if the variable 'Average Area of Number of Bedroom' is insignificant to our model. Our null hypothesis will be our reduced model which will include the variables: 'Average Area Income', 'Average Area House Age', 'Average Area Number of Rooms', and 'Area Population' versus the entire full model with all the predictors. Looking at the general linear F-Test, we can conclude that the reduced model was statistically better thus showing that the variable 'Average Area of Number of Bedroom' is insignificant. We had a p-value greater than our alpha=0.05, thus failing to reject our null hypothesis and the results match our T-test from question 1.



```
Analysis of Variance Table

Model 1: Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms +
    Area.Population
Model 2: Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms +
    Area.Population + Avg..Area.Number.of.Bedrooms
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1   4995 5.1115e+13
2   4994 5.1099e+13  1 1.6288e+10 1.5919 0.2071
```

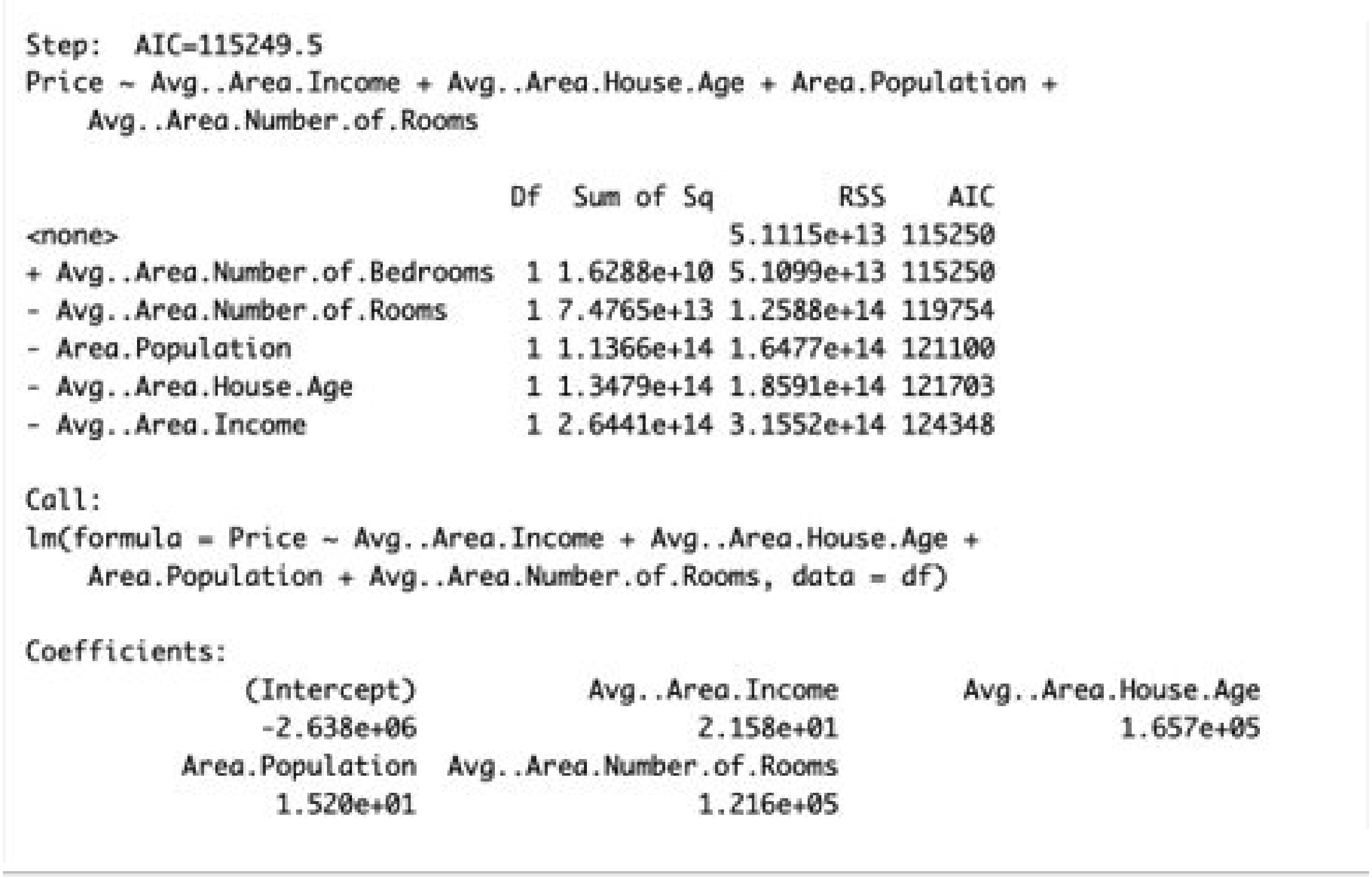
Now to answer question 2 to test if an interaction model would be better to predict housing prices, we used another linear F-Test to see if a reduced model with all variables except for 'Average Area of Number of Bedroom' and the full model that includes interaction effects with all variables. The Analysis of Variance table returned an F-Stat for 0.4561 and p-value of 0.8411 and conclude that we fail to reject the null hypothesis thus making our reduced model the better statistically significant model. Hence, the interactions between our numeric variables are insignificant and do not improve our model. There is a disclaimer here, where we tested all the interactions variables at once and the general linear F-test concluded that having all interactions were insignificant. However, it is worthy to mention that maybe a pair or a single interaction may be useful to the model and will require further testing. There is another method where we can use step wise regression.



```
Analysis of Variance Table

Model 1: Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms +
    Area.Population
Model 2: Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms +
    Area.Population + Avg..Area.Income * Avg..Area.House.Age +
    Avg..Area.Income * Avg..Area.Number.of.Rooms + Avg..Area.Income *
    Area.Population + Avg..Area.House.Age * Avg..Area.Number.of.Rooms +
    Avg..Area.House.Age * Area.Population + Avg..Area.Number.of.Rooms *
    Area.Population
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1   4995 5.1115e+13
2   4989 5.1087e+13  6 2.8021e+10 0.4561 0.8411
```

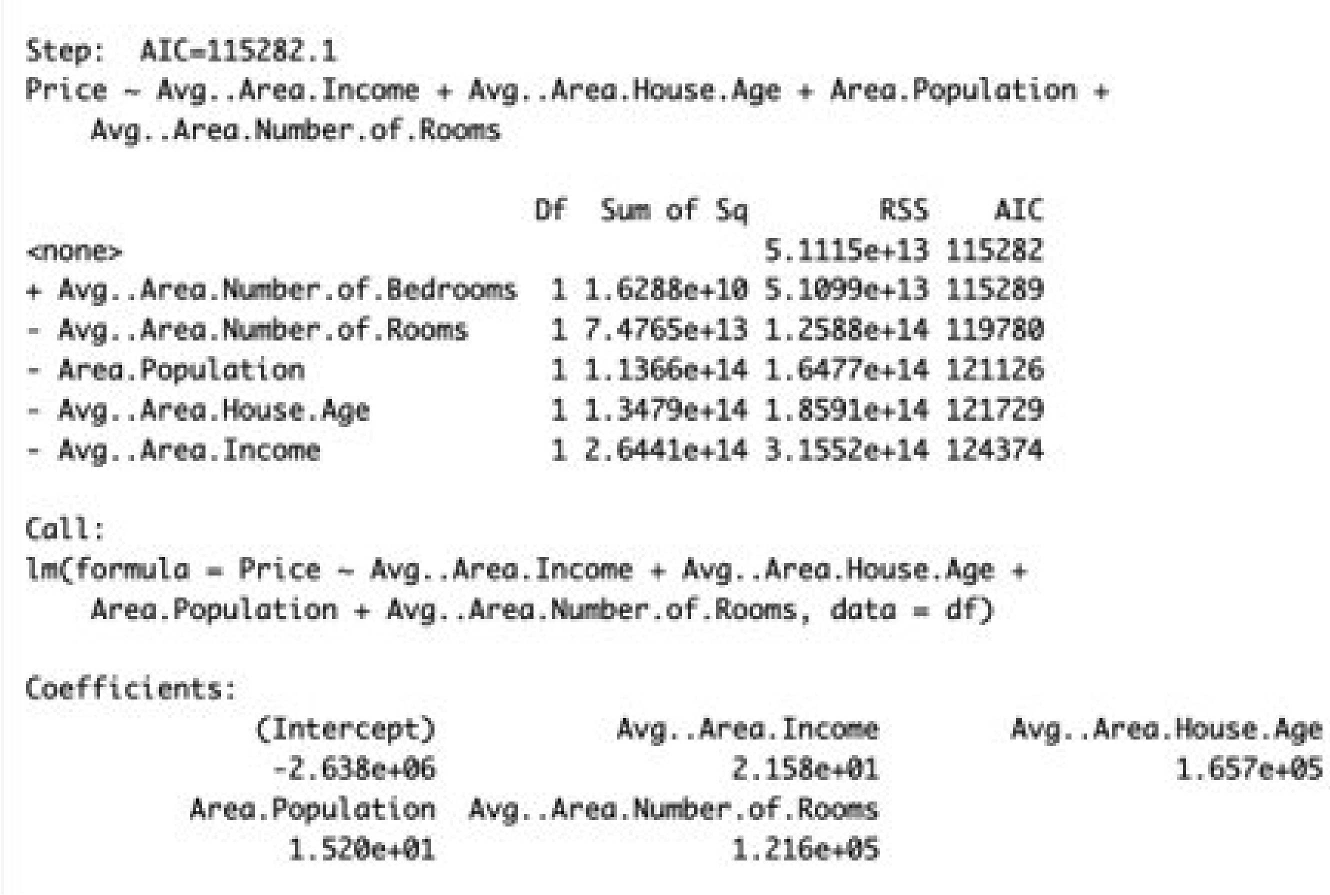
Next, we used a step-wise regression using AIC and BIC as our metrics to test what subset of variables would make the best models. The only difference between the two metrics mention is their penalty term which AIC uses the constant 2 as a penalty and BIC using $\log(n)$ as the penalty term. The results of the step wise regression using both metrics returned similar results where both metrics had close values. The values of BIC were greater implying it did slightly worse but it was hardly noticeable. It is worth mentioning that BIC works better with smaller models. Using step-wise regression the predictor ’Average Area Income’ was the most significant predictor and our final model was the same model that the t-test and overall f-test concluded to be the most significant.

AIC Score



```
Step:  AIC=115249.5
Price ~ Avg..Area.Income + Avg..Area.House.Age + Area.Population +
    Avg..Area.Number.of.Rooms

                              Df  Sum of Sq        RSS    AIC
<none>                                      5.1115e+13 115250
+ Avg..Area.Number.of.Bedrooms  1 1.6288e+10 5.1099e+13 115250
- Avg..Area.Number.of.Rooms     1 7.4765e+13 1.2588e+14 119754
- Area.Population               1 1.1366e+14 1.6477e+14 121100
- Avg..Area.House.Age           1 1.3479e+14 1.8591e+14 121703
- Avg..Area.Income              1 2.6441e+14 3.1552e+14 124348

Call:
lm(formula = Price ~ Avg..Area.Income + Avg..Area.House.Age +
    Area.Population + Avg..Area.Number.of.Rooms, data = df)

Coefficients:
              (Intercept)           Avg..Area.Income        Avg..Area.House.Age
               -2.638e+06                  2.158e+01                  1.657e+05
          Area.Population  Avg..Area.Number.of.Rooms
                1.520e+01                  1.216e+05
```

BIC Score



```
Step:  AIC=115282.1
Price ~ Avg..Area.Income + Avg..Area.House.Age + Area.Population +
    Avg..Area.Number.of.Rooms

                             Df  Sum of Sq        RSS    AIC
<none>                                     5.1115e+13 115282
+ Avg..Area.Number.of.Bedrooms  1 1.6288e+10 5.1099e+13 115289
- Avg..Area.Number.of.Rooms     1 7.4765e+13 1.2588e+14 119780
- Area.Population               1 1.1366e+14 1.6477e+14 121126
- Avg..Area.House.Age           1 1.3479e+14 1.8591e+14 121729
- Avg..Area.Income              1 2.6441e+14 3.1552e+14 124374

Call:
lm(formula = Price ~ Avg..Area.Income + Avg..Area.House.Age +
    Area.Population + Avg..Area.Number.of.Rooms, data = df)

Coefficients:
              (Intercept)           Avg..Area.Income        Avg..Area.House.Age
               -2.638e+06                  2.158e+01                  1.657e+05
          Area.Population  Avg..Area.Number.of.Rooms
                1.520e+01                  1.216e+05
```

### 4.5 Q3. What is the average house price in the united states when the average area income is $50,000, average area house age is 5, an area population of 35,000 people, and average area number of rooms is 6?

We have found the best subset of predictors and satisfied all of our assumptions about the residuals along with linearity, so we can now use this model to answer further research questions. Since we want to find the average house price given certain attributes of a house, this is considered to be a confidence problem, where we use the predicted linear regression model to find the solution and standard error to find our interval. Given a house which consist of average area income of $50,000, average house age is 5, an area population of 35,000 people, and average area number of rooms is 6, this describes a typical American living in a city. Our predicted linear regression model returned a house price worth $531,319, which seems about right. Our 95% confidence interval returned a house price between $[\$524,410 : \$538,227]$. Now, how does this interval or price compare to your neighborhood.

## 5 Conclusion

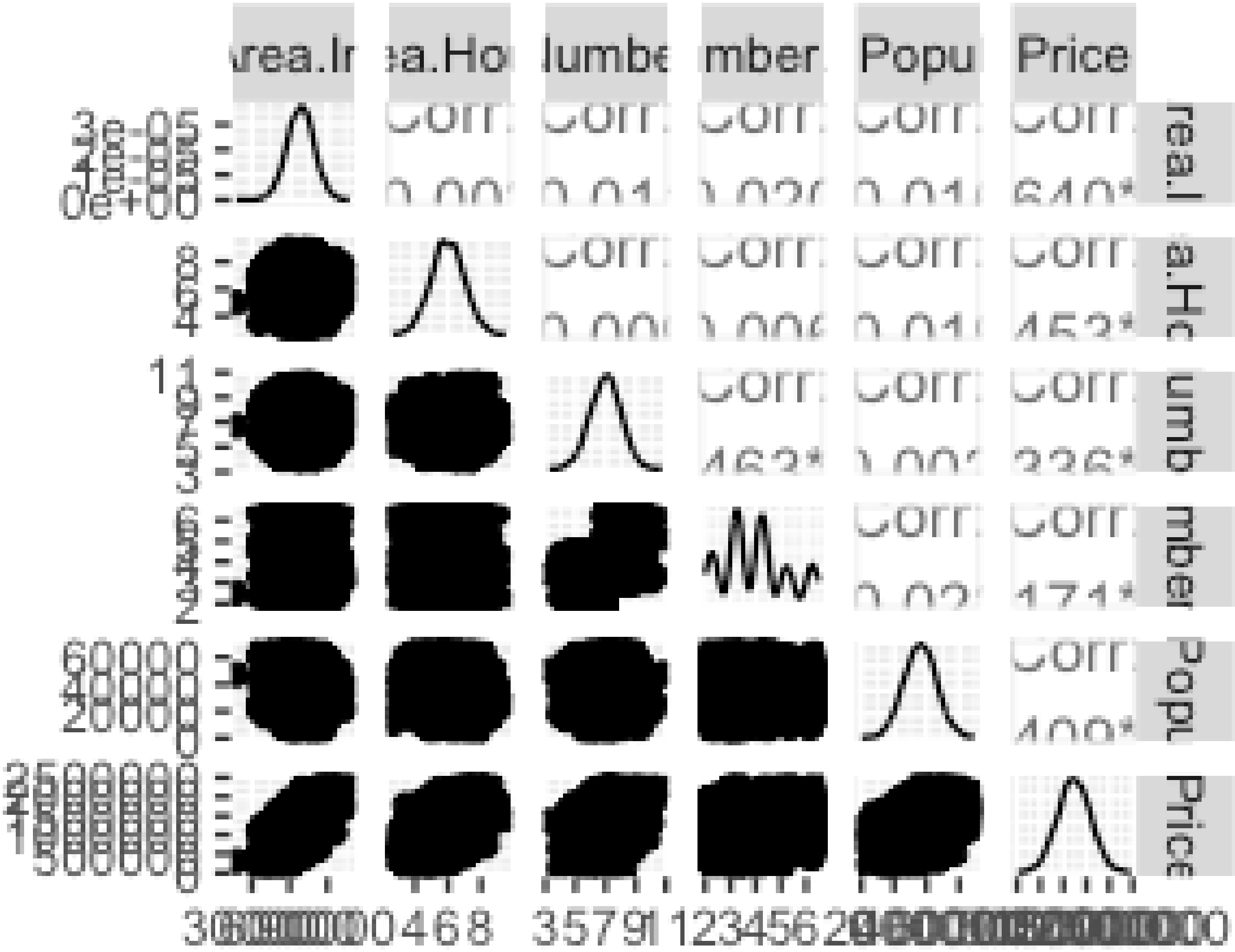
First we performed exploratory data analysis to check for any inconstancy in our data and cleaned it up. After this step we performed several data visualizations to get any further insights and plotted a correlation matrix to check how are the variables in the data set are related. This step is crucial in determining what predictors to use but further statistical test must be used to accurately identify them. We fitted a model using all predictors and performed residual analysis to check our four assumptions which did not lead to any transformations performed on our model. Then we conducted a t-test on the slope parameters to look for the most significant predictors. We found only 1 insignificant predictor and conducting an overall F-test resulted it the same conclusion along with the t-test. We tested if the interactions between the variables could improve our model and the general linear F-test resulted that the interactions were insignificant. After finding the best model we used it to answer a prediction problem where the goal was to find the average housing price given certain house attributes. Overall we had a good model which satisfied all assumption and resulted in perfectly linear plot.

If we had more time to work on our project, we could have improved our model by including addresses and grouping different zip codes to better define pricing for different local areas. It is well-known that the price of housing is complex and influenced by local factors. Therefore, we could have included these variables through classification and running different subset models to obtain more accurate representations of local prices, rather than relying on a more general approach based on the US housing market average prices. It is important to have a macro approach, but a more dense micro approach is needed to provide the reader with the best price scenarios for making better-informed decisions. Additionally, we could have incorporated external variables in the market, such as measuring inflation on the dollar, that have an indirect impact on price. In summary, this model provides an overall average price, but it is not a definitive or robust model for accurately predicting housing prices. The hope of this report is to inform the reader about the average housing prices in the United States through the lens of linear regression.

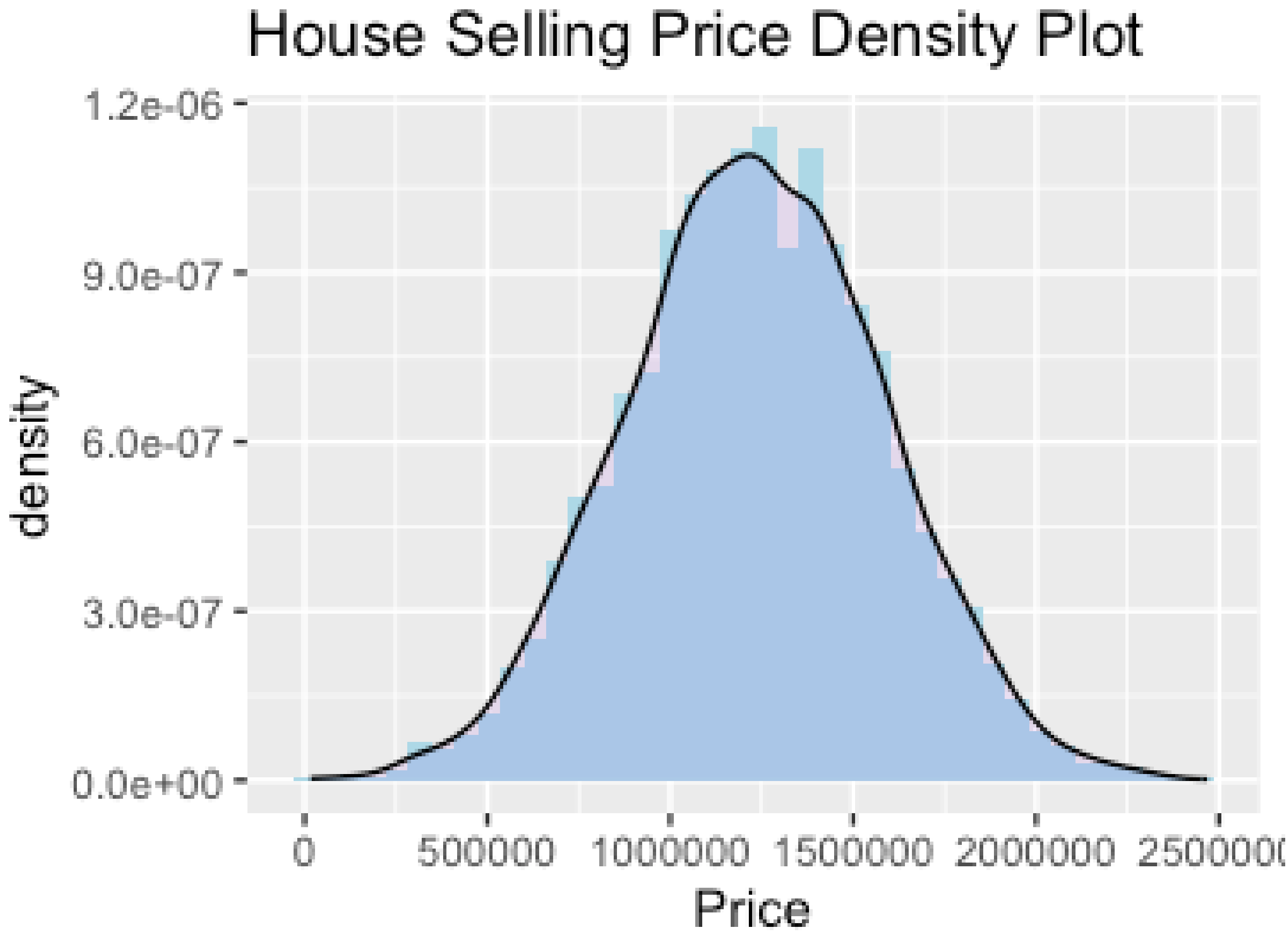
# 6 Appendix

## 6.1 Visualizations

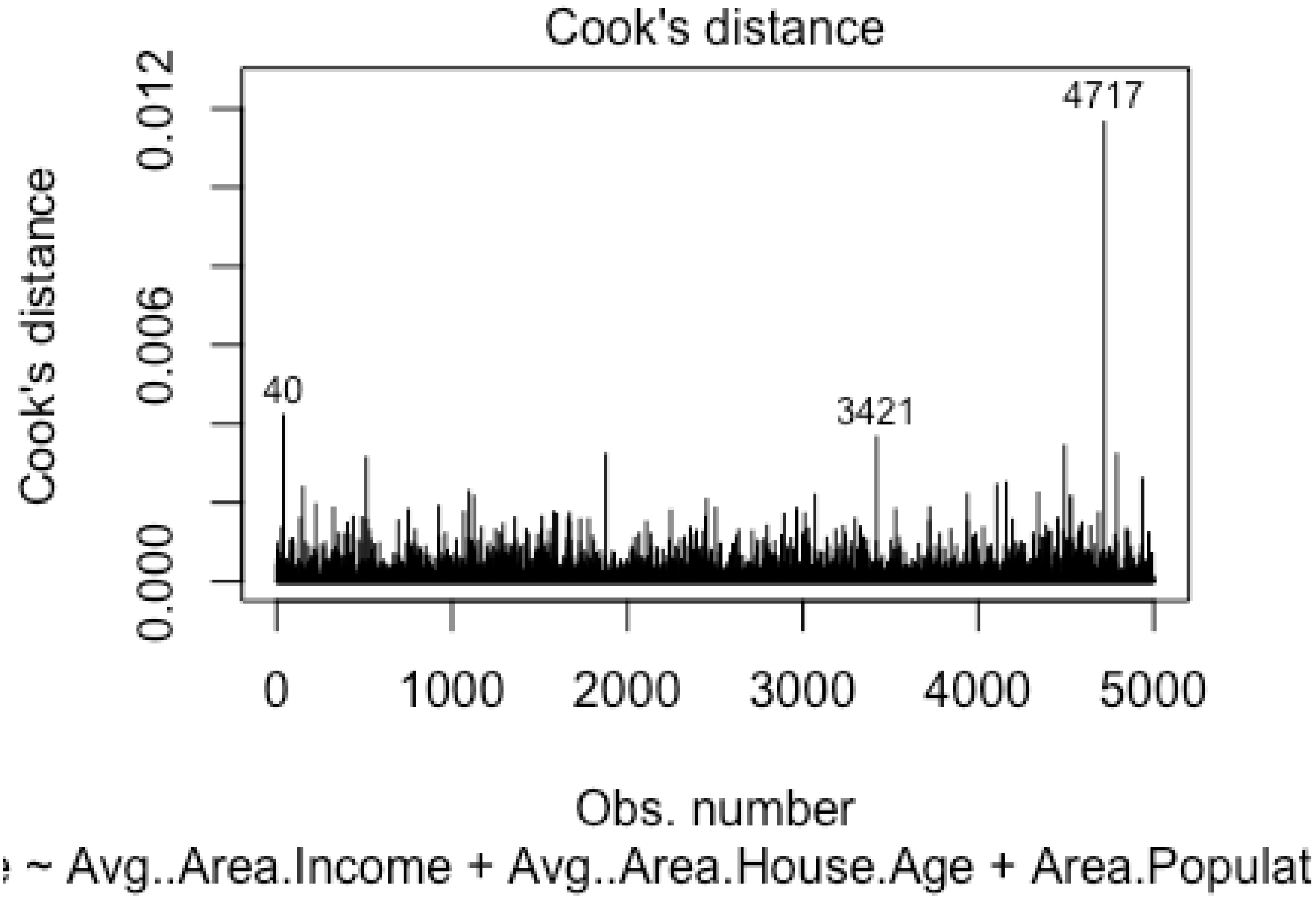
Scatter plots



Price Density Histogram



Cooks Distance (test for influential points)



# References

[1] @websiteKaggle, author = Vyas, title = *USA_HOUSING*, url = https://www.kaggle.com/datasets/vedavyasv/usa-housing, urldate = 2023-04-30, originalyear = 2017-11-19

## 6.2 R Code

```
library('tidyverse')
library('MASS')
library('glmnet')
library('ggplot2')
library('corrplot')
library('GGally')
library('ISLR')
library('skimr')


## ------------------------------------------------------------------------
# Load the data
```

```
df = read.csv('~/Stat510/Stat510_S23/datasets/USA_Housing.csv')
df = data.frame(df)
glimpse(df)


## ----------------------------------------------------------------------------
# data dimensions
dim(df)


## ----------------------------------------------------------------------------
# check for missing values in the columns
colSums(is.na(df))

## ----------------------------------------------------------------------------
# check for duplicates
sum(duplicated(df))


## ----------------------------------------------------------------------------
# statistical summary using skim function from skimr
skim(df)

## ----------------------------------------------------------------------------
# convert categorical variables into factors
factor_names = c('Address')
df = df |> mutate_at(factor_names, as.factor)


## ----------------------------------------------------------------------------
# check column data types
str(df)


## ----------------------------------------------------------------------------
# correlation matrix plot

# color palette
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA

# extract numeric columns only
numeric_cols <- sapply(df, is.numeric)
df_numeric <- df[, numeric_cols]

corr_matrix <- cor(df_numeric)
corrplot(corr_matrix, method = 'color', tl.cex = 0.5, title = "Correlation□Ma
        mar=c(0,0,1,0), addCoef.col = 'grey50')


## ----------------------------------------------------------------------------
# scatter plot matrix
ggpairs(df_numeric)
```

```
## -----------------------------------------------------------------
# Histogram on Average Area Income
ggplot(df, aes(x = Avg..Area.Income)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40, fill = ”lightblue”)
  geom_density(alpha = 0.1, fill = ”lightgreen”) +
  labs(title=”Average␣Area␣Income␣Density␣Plot”,x=”Average␣Area␣Income”)


## -----------------------------------------------------------------
# Histogram on Average Area House Age
ggplot(df, aes(x = Avg..Area.House.Age)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40, fill = ”lightblue”)
  geom_density(alpha = 0.1, fill = ”lightgreen”) +
  labs(title=”Average␣Area␣House␣Age␣Density␣Plot”,x=”Average␣Area␣House␣Age”)


## -----------------------------------------------------------------
# Histogram on Area Population
ggplot(df, aes(x = Area.Population)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40, fill = ”lightblue”)
  geom_density(alpha = 0.1, fill = ”purple”) +
  labs(title=”Area␣Population␣Density␣Plot”,x=”Area␣Population”)


## -----------------------------------------------------------------
# Histogram on Average Area Number of Bedrooms
ggplot(df, aes(x = Avg..Area.Number.of.Bedrooms)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40, fill = ”lightblue”)
  geom_density(alpha = 0.1, fill = ”lightgreen”) +
  labs(title=”Average␣Area␣Number␣of␣Bedrooms␣Density␣Plot”,x=”Average␣Area␣Nu


## -----------------------------------------------------------------
# Histogram on Average Area Number of Rooms
ggplot(df, aes(x = Avg..Area.Number.of.Rooms)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40, fill = ”lightblue”)
  geom_density(alpha = 0.1, fill = ”purple”) +
  labs(title=”Average␣Area␣Number␣of␣Rooms␣Density␣Plot”,x=”Average␣Area␣Numbe

## -----------------------------------------------------------------
# Histogram on Average Area Number of Rooms
ggplot(df, aes(x = Price)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40, fill = ”lightblue”)
  geom_density(alpha = 0.1, fill = ”purple”) +
  labs(title=”House␣Selling␣Price␣Density␣Plot”,x=”Price”)


## -----------------------------------------------------------------
# remove the categorical column since each address is unique
df = df |> dplyr::select(-Address)
```

```
## ----------------------------------------------------------------------------
# split the data
# split train and test sets to a 80/20 split
n = nrow(df)
prop = .80
set.seed(1)
train_id = sample(1:n, size = round(n*prop), replace = FALSE)
test_id = (1:n)[-which(1:n %in% train_id)]
train_set = df[train_id, ]
test_set = df[test_id, ]


## ----------------------------------------------------------------------------
# Fit a Linear Regression Model with all predictors
linear.fit = lm(Price ~ ., data = df)
summary(linear.fit)


## ----------------------------------------------------------------------------
# residual vs fit plot for our linear model
residuals = linear.fit$residuals
fitted_values = linear.fit$fitted.values
plot(fitted_values,residuals, main = ’Residual␣Vs␣Fit␣Plot␣(Linear␣Fit)’,
     xlab = ’Fitted␣Values’, ylab = ’Residuals’, col = ’chocolate1’)
abline(0,0, lty=3)


## ----------------------------------------------------------------------------
# QQ-plot on residuals for our linear model
qqnorm(residuals, pch = 1, frame = TRUE)
qqline(residuals, col = ”steelblue”, lwd = 2)


## ----------------------------------------------------------------------------
# Fitting the Linear model with only significant predictors
# we got rid of the average area number of bedrooms
linear.fit2 = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Nu
                  Area.Population, data = df)
summary(linear.fit2)


## ----------------------------------------------------------------------------
# residual vs fit plot for our linear model with significant predictors
residuals = linear.fit2$residuals
fitted_values = linear.fit2$fitted.values
plot(fitted_values,residuals, main = ’Residual␣Vs␣Fit␣Plot␣(Linear␣Fit)’,
     xlab = ’Fitted␣Values’, ylab = ’Residuals’, col = ’chocolate1’)
abline(0,0, lty=3)


## ----------------------------------------------------------------------------
# Residual versus Average area Income
plot(df$Avg..Area.Income,residuals, main = ’Residual␣Vs␣Avg.␣Area␣Income’,
     xlab = 'Avg.□Area□Income', ylab = 'Residuals', col = 'chocolate2')
abline(0,0, lty=3)


## ----------------------------------------------------------------------
# Resudlas versus Average Area House Age
plot(df$Avg..Area.House.Age,residuals, main = 'Residual□Vs□Avg.□Area□House□Age
     xlab = 'Avg.□Area□House□Age', ylab = 'Residuals', col = 'chocolate2')
abline(0,0, lty=3)


## ----------------------------------------------------------------------
# Residulas versus Area Population
plot(df$Area.Population,residuals, main = 'Residual□Vs□Area□Population',
     xlab = 'Area□Population', ylab = 'Residuals', col = 'chocolate2')
abline(0,0, lty=3)


## ----------------------------------------------------------------------
# Residulas versus Average Area Number of Bedrooms
plot(df$Avg..Area.Number.of.Bedrooms,residuals, main = 'Residual□Vs□Avg.□Area□
     xlab = 'Avg.□Area□Number□of□Bedrooms', ylab = 'Residuals', col = 'chocola
abline(0,0, lty=3)

## ----------------------------------------------------------------------
# QQ-plot on residuals for our linear model
qqnorm(residuals, pch = 1, frame = TRUE)
qqline(residuals, col = "steelblue", lwd = 2)

## ----------------------------------------------------------------------
# test for normality on residuals for our linear model
shapiro.test(residuals)


## ----------------------------------------------------------------------
# Transformation Attempt
linear.transform = lm(Price ~ log(Avg..Area.Income) + log(Avg..Area.House.Age)
                      log(Avg..Area.Number.of.Rooms) + log(Area.Population), d
summary(linear.transform)

## ----------------------------------------------------------------------
# residual vs fit plot for our Transformed linear model
residuals = linear.transform$residuals
fitted_values = linear.transform$fitted.values
plot(fitted_values,residuals, main = 'Residual□Vs□Fit□Plot□(Linear□Fit)',
     xlab = 'Fitted□Values', ylab = 'Residuals', col = 'chocolate1')
abline(0,0, lty=3)

## ----------------------------------------------------------------------
# QQ-plot on residuals for our Transformed linear model
qqnorm(residuals, pch = 1, frame = TRUE)
qqline(residuals, col = "steelblue", lwd = 2)
```

```
## ---------------------------------------------------------------------
# test for normality on residuals for our transformed linear model
shapiro.test(residuals)


## ---------------------------------------------------------------------
# Reduced Linear Model
reduced.model = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.
                   Area.Population, data = df)

# Full Linear Model
full.model = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Num
                   Area.Population + Avg..Area.Number.of.Bedrooms, data = df)

# General Linear F-Test
anova(reduced.model, full.model)


## ---------------------------------------------------------------------
# Reduced Linear Model
reduced.model = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.
                   Area.Population, data = df)

# Full Linear Model
full.model = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Num
                   Area.Population + Avg..Area.Income*Avg..Area.House.Age +
                   Avg..Area.Income*Avg..Area.Number.of.Rooms +
                   Avg..Area.Income*Area.Population + Avg..Area.House.Age*Avg..
                   Avg..Area.House.Age*Area.Population +
                   Avg..Area.Number.of.Rooms*Area.Population, data = df)

# Summary of Full Model
summary(full.model)

## ---------------------------------------------------------------------
# General Linear F-Test
anova(reduced.model, full.model)


## ---------------------------------------------------------------------
# Step-Wise Regression using AIC
mod0 = lm(Price ~ 1, data = df)
mod.upper = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Numb
                   Area.Population + Avg..Area.Number.of.Bedrooms, data = df)
step(mod0, scope = list(lower = mod0, upper = mod.upper), k = 2)


## ---------------------------------------------------------------------
# Step-Wise Regression using AIC
mod0 = lm(Price ~ 1, data = df)
mod.upper = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Numb
                   Area.Population + Avg..Area.Number.of.Bedrooms, data = df)
step(mod0, scope = list(lower = mod0, upper = mod.upper), k = log(5000))
```

```
## ------------------------------------------------------------------------
# The best model to answer research questions
best.model = lm(formula = Price ~ Avg..Area.Income + Avg..Area.House.Age +
    Area.Population + Avg..Area.Number.of.Rooms, data = df)
summary(best.model)

## ------------------------------------------------------------------------
# 95\% confidence intervals on the coefficients of the best model
confint(best.model, level = 0.95)

## ------------------------------------------------------------------------
# 95\% average Confidence prediction
new = data.frame(Avg..Area.Income = 50000, Avg..Area.House.Age = 5,
                 Area.Population = 35000, Avg..Area.Number.of.Rooms = 6)
ans = predict(best.model, new, se.fit = TRUE, interval = ”confidence”, level =
ans

## ------------------------------------------------------------------------
# check for potential influential points
plot(best.model, which = 4)
abline(h = 0.5, lty = 2)
```