



**Table of Contents**

[**Purpose 3**](#_heading=h.30j0zll)

[**Our Data 3**](#_heading=h.3znysh7)

[**Background on Dataset & Variables 3**](#_heading=h.tyjcwt)

[**Part 1 - Model Estimation and Interpretation 4**](#_heading=h.1t3h5sf)

[**1a. Fitting a regression model estimating sales price using beds, baths, and garage size as predicting variables 4**](#_heading=h.2s8eyo1)

[**1b. Interpretation of Coeﬀicients 4**](#_heading=h.3rdcrjn)

[**1c. Interpretation of Adjusted R-Squared = 0.54 5**](#_heading=h.lnxbz9)

[**Part 2 - Prediction 5**](#_heading=h.1ksv4uv)

[**2a. Predicting the house sales price for a house with 3 bedrooms, 3 bathrooms, and a 2-car garage 5**](#_heading=h.2jxsxqh)

[**2b. Calculating the 95% confidence interval 5**](#_heading=h.3j2qqm3)

[**2c. Calculating the 95% prediction interval 5**](#_heading=h.4i7ojhp)

[**Part 3 - Hypothesis Testing 6**](#_heading=h.1ci93xb)

[**3a. Checking the significance for each individual partial slope (independent variable) 6**](#_heading=h.2bn6wsx)

[**3b. Conducting an F-test to check overall model significance 7**](#_heading=h.3as4poj)

[**3c. Conducting Partial F tests 7**](#_heading=h.49x2ik5)

[**Part 4 - Multicollinearity 8**](#_heading=h.147n2zr)

[**4a. Creating scatterplots and correlation matrices 9**](#_heading=h.23ckvvd)

[**4b. Removing Two Strongly Correlated Variables 11**](#_heading=h.32hioqz)

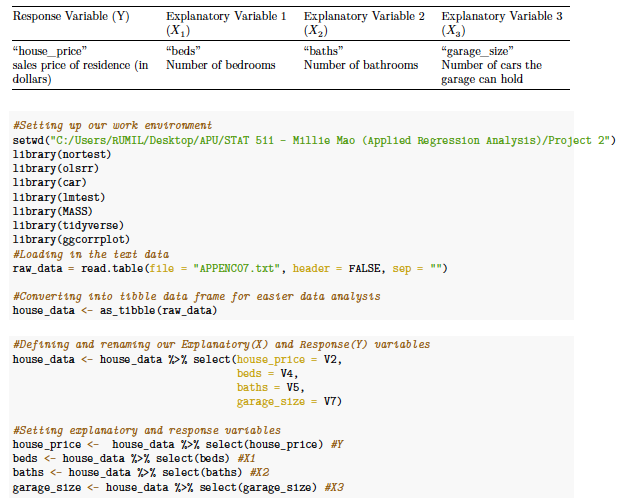
# Purpose

We are conducting a multiple linear regression from the Real Estate Sales (APPENC07) dataset to analyze the relationship of the given features, bedrooms, bathrooms, and garage size, with the outcome variable, house sales price in a Midwestern city.

# Our Data

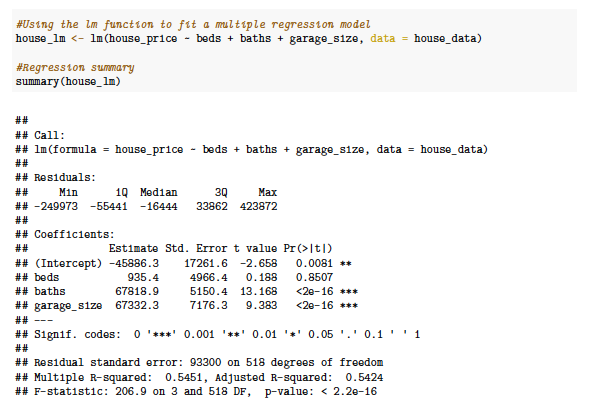
## Background on Dataset & Variables

Our dataset consists of ***522 total transactions*** from home sales during the year 2002.



# Part 1 - Model Estimation and Interpretation

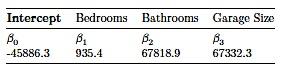
## 1a. Fitting a regression model estimating sales price using beds, baths, and garage size as predicting variables

****

## 1b. Interpretation of Coeﬀicients

***Intercept & Partial Slopes***

From summarizing our multiple linear regression model we can see:



and the estimated regression equation to be:

**𝑌̂ = −45886.3 + 935.4𝑋 + 67818.9𝑋 + 67332.3𝑋**

The partial slopes in our summary indicate that when any one of the partial slopes **Increase by 1 unit**

and **other explanatory variables held constant and unchanged** we can expect:

* While holding our other explanatory variables Bathrooms and Garage Size constant and unchanged, when **Bedrooms** increase by 1 unit, we can expect our **house sales price** to increase by **roughly $935.40**
* While holding our other explanatory variables Bedrooms and Garage Size constant and unchanged, when **Bathrooms** increases by 1 unit, we can expect our **house sales price** to increase by **roughly $67,818.90.**
* While holding our other explanatory variables Bedrooms and Bathrooms constant and unchanged, when **Garage size** increases by 1 unit, we can expect our **house sales price** to increase by **roughly $67,332.30.**

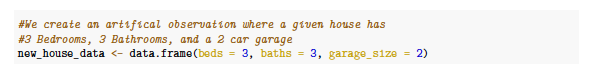
## 1c. Interpretation of Adjusted R-Squared = 0.54

Adjusted R squared value, similar to the R square value, tells us how much of the variability in our model is explained by our predictor variables, while also penalizing redundant or otherwise useless predictor variables, helping us to resist urges of adding too many variables into our model.

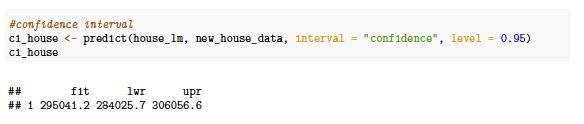
In this case our adjusted R² of 0.54 tells us that about **54% of the variation in our response variable is explained by our 3 explanatory variables**.

# Part 2 - Prediction

## 2a. Predicting the house sales price for a house with 3 bedrooms, 3 bathrooms, and a 2-car garage



## 2b. Calculating the 95% confidence interval



**Interpretation**

This 95% confidence interval, when Bedrooms = 3, Bathrooms = 3, and Garage Size = 2, is from 74.84094 to 79.70906.

When Bedrooms = 3, Bathrooms = 3, and Garage Size = 2, with 95% confidence we can expect our confidence interval to capture the average of house sales prices (response variable).

## 2c. Calculating the 95% prediction interval

****

**Interpretation**

From the results we can predict with 95% confidence that when there are 3 bedrooms, 3 bathrooms, and a garage that can hold 2 cars, the predicted house sales price will fall somewhere between **111,422 to 478,660 dollars.**

# Part 3 - Hypothesis Testing

## 3a. Checking the significance for each individual partial slope (independent variable)

**Using a significance level of** 𝛼 = 0.05

Null Hypothesis: 𝐻0: 𝛽𝑗 = 0 (slopes are showing no change), 𝑋𝑗 **is not** linearly associated with Y, therefore the partial slope **is not significant.**

Alternative Hypothesis: 𝐻1: 𝛽𝑗 ≠ 0 (slopes are showing change), **𝑋𝑗** is linearly associated with Y, therefore the partial slope **is significant.**

Table 3: Table Representation of Hypothesis Testing



Bedrooms (𝑋1) Bathrooms (𝑋2) Garage Size (𝑋3) 0.8507 **>** 𝛼 = 0.05 <2e-16 **<** 𝛼 = 0.05 <2e-16 **<** 𝛼 = 0.05

Fail to reject 𝐻0 Reject 𝐻0 Reject 𝐻0

Not Significant Significant Significant



**Bedroom variable:**

The p-value of Bedroom is 0.8507 and is greater than our 𝛼 (accepted error) of 0.05, so we fail to reject our NULL hypothesis and must conclude with our NULL hypothesis. Stating that our partial slope, Bedrooms, does not show overall significance in our model.

**Bathroom & Garage Size variables:**

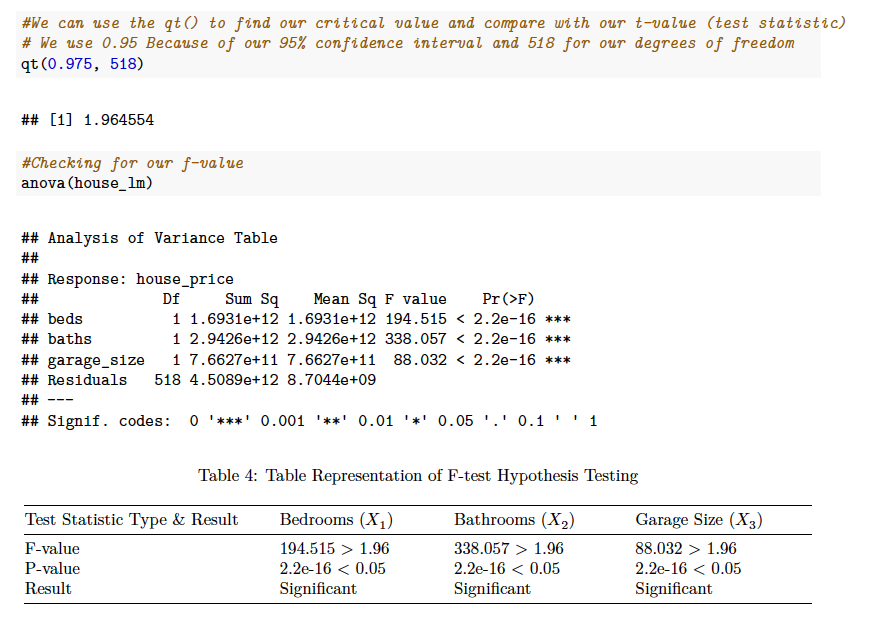
On the other hand, because the p-value of Bathroom and Garage size are both <2e-16 and are incredibly smaller than our (accepted error) of 0.05, so we reject our NULL hypothesis and conclude with our alternative hypothesis. Our alternative hypothesis states that our partial slopes, Bathroom and Garage Size, shows overall significance in our model.

## 3b. Conducting an F-test to check overall model significance

**Using a significance level of** 𝛼 = 0.05

**Null Hypothesis:** 𝐻0: 𝛽1 = 𝛽2 = 0 (**No** partial slopes are significant). Shows no change, therefore does **not show** overall model significance.

**Alternative Hypothesis**: 𝐻1: 𝛽1 = 𝛽2 = ≠ 0 (**At least one** partial slope is significant). Shows change, therefore **showing** overall model significance.



Our p-value of < 2.2e-16 being less than our alpha and our F values being larger than our critical value tells us we can **reject** our NULL hypothesis and conclude with our alternative hypothesis, that **at least one** of our predictor variables **shows** overall model significance.

## 3c. Conducting Partial F tests

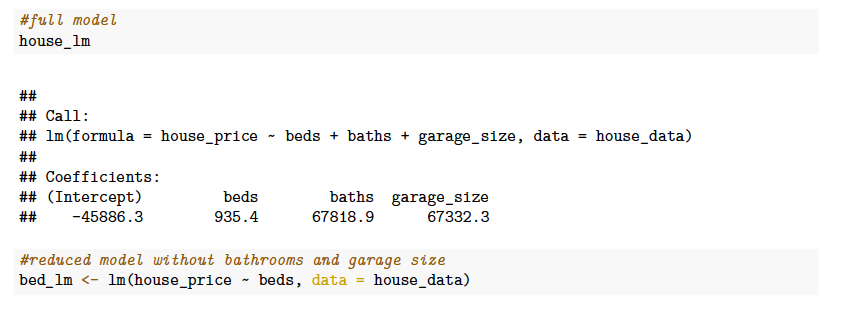
**Which variable is actually contributing?**

Conducting partial F tests is important to see if the number of bathrooms (X2) and garage size(X3) are jointly significant.

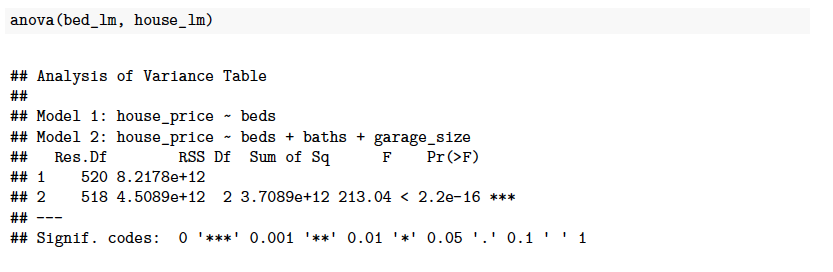
**Using a significance level of** 𝛼 =0.05

**Null Hypothesis:** 𝐻0: **There is no change** when adding certain predictors to the significance of our model

**Alternative Hypothesis:** 𝐻1: **There is change** when adding certain predictors towards the significance of our model.



We now ***compare*** our reduced model with our complete model

Since the p-value is **2.2e-16** is less than our significance level of **0.05** we see that bathroom and garage size are both jointly significant and therefore we can reject the null hypothesis, indicating there is significance in keeping both bathroom and garage size in our model.

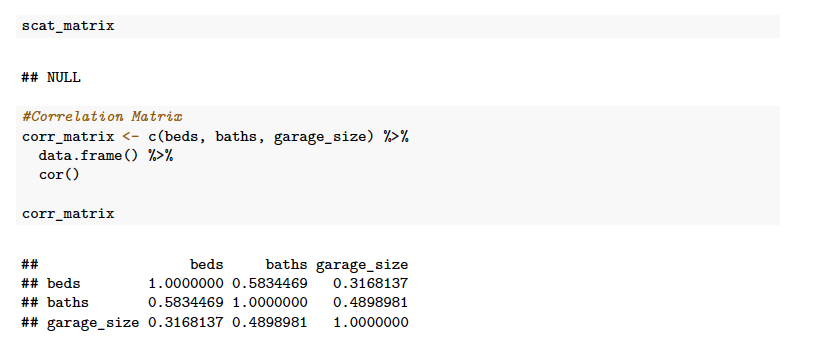
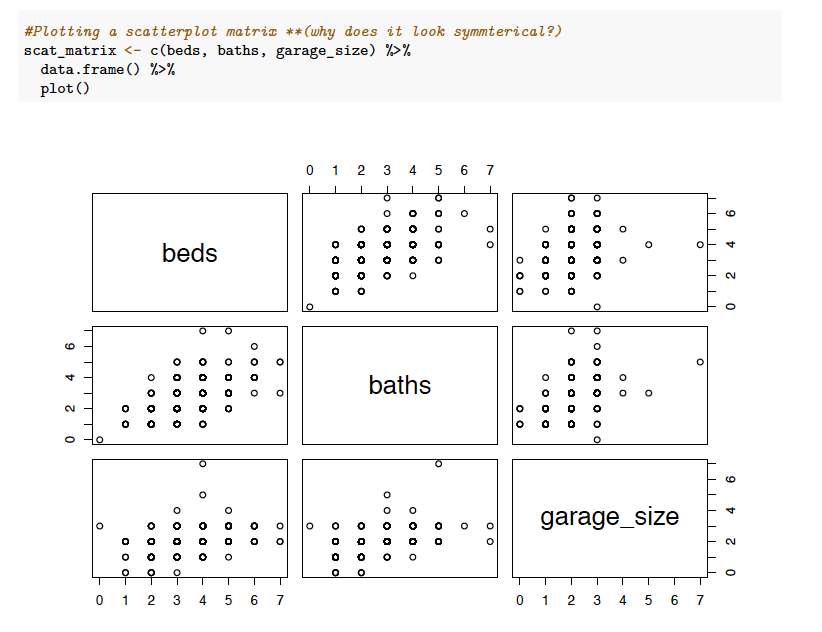
In effect, we are concluding that bathroom and garage size are predictors that do contribute information in the prediction of house sales price and therefore should be retained in the model.

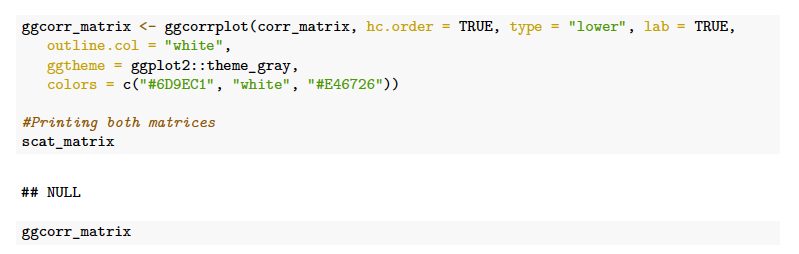
# Part 4 - Multicollinearity

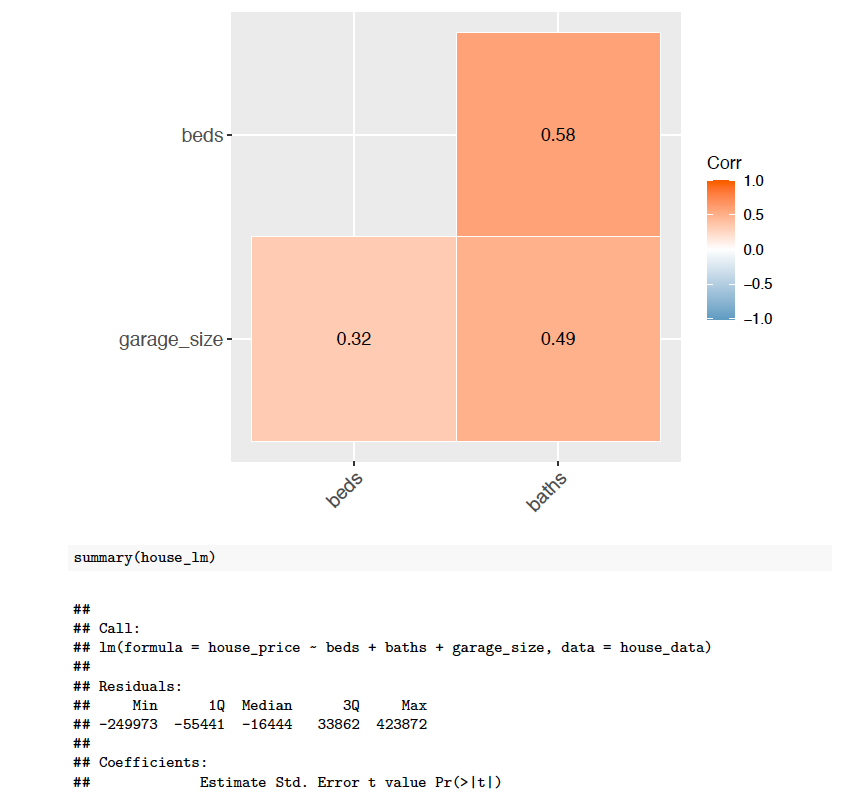
**Why bother with multicollinearity?**

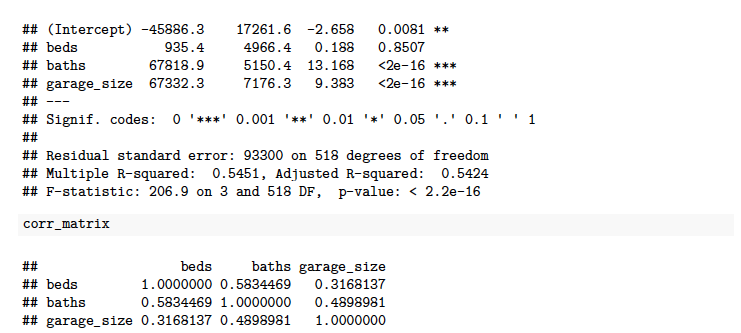
Having multicollinearity is problematic because by having multiple correlated predictor variables, it becomes harder for our model to attribute significance to our predictor variables. It creates redundant and duplicate information, thereby negatively affecting the results of our regression model.

## 4a. Creating scatterplots and correlation matrices





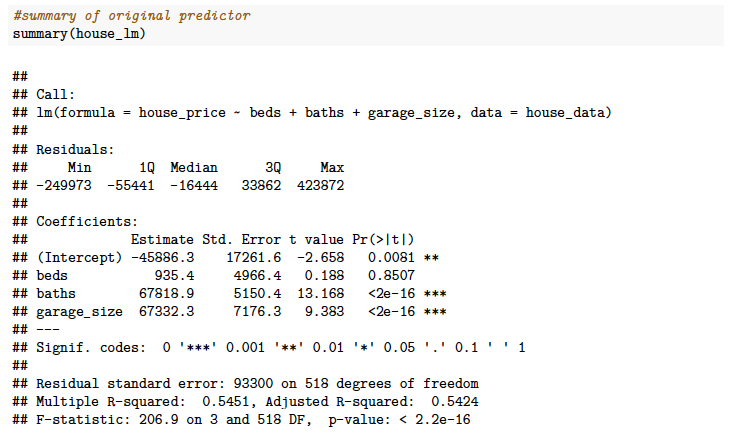




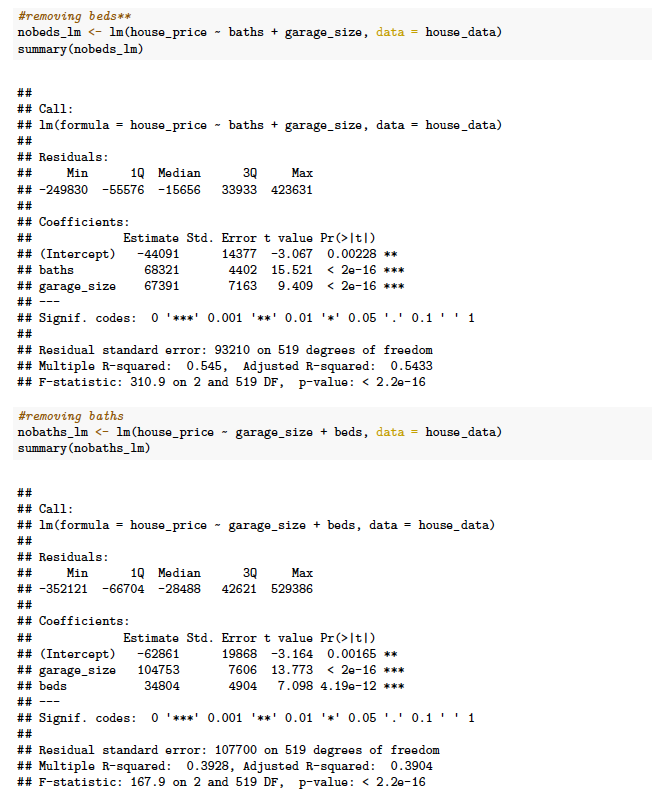
## 4b. Removing Two Strongly Correlated Variables

A way to combat this is by removing a highly correlated predictor. From the correlation matrix and by looking at our correlation coeﬀicient, we can see moderately positive relationship between bedrooms and bathrooms which might be worth further investigating.

We know there is some kind of multicollinearity issue with bed and baths, I am more interested in beds than baths. So I can remove baths from our model thereby correcting our multicollinearity issue.



We notice bedrooms is not a significant variable from looking at the p-value, when in fact, the reality is **it should** be significant. Knowing this, we can check to see how well our model performs when removing bathroom since there is a multicollinearity issue.



In both cases we can see that by either removing bedroom or bathroom in our model, the predictors still remain significant but more importantly we can now see that bedrooms is in fact a significant predictor when removing the bathroom variable in our model concluding that we have addressed our issue of multicollinearity.