**`ISOM 670 Stuk Student ID Number**: **2234611**

**Final Quiz, August 2018**

HONOR CODE

The Goizueta Business School Honor Code is the standard of professional behavior on this exam. When you have completed your exam, please read the following pledge and add your signature if you have complied with the Honor Code.

*I pledge that I have neither given nor received any unauthorized assistance on this exam, and that any violations of the Honor Code by others that I have observed or otherwise become aware of will be reported by me to the Honor Council.*

**Type Name to Confirm: Carl Xi**

INSTRUCTIONS — READ CAREFULLY

During the Decision Analysis exam **all** discussion related to the exam with anyone (other than the professor) is prohibited. Even after you have turned in your own work, you still may not discuss any particulars of the exam until we have indicated that the entire class has submitted their work. Please take great care not to carelessly or inadvertently cause an Honor Code violation.)

**Exam Mechanics:**

Use only your own notes and exam prep materials. Sharing materials with other students during the exam period is not permitted. You may use anything posted in our course conference, whether you downloaded it before the exam or not.

Computers are permitted throughout and are necessary for some parts. You are not required to use computers if there is another way to get to an answer.

We have not provided space in the exam booklet itself for you to show your work. Please adjust the spacing accordingly when you create the printed version that you will be turning in.

**Transfer your answers** to the front Answer Sheet when indicated. Failure to do this may cost you some points. Of course, your work pages will contain any long answers & required explanations that might accompany the short answers. They will also show your assumptions and how you got your answers. (Note: we recommend that you support all answers by showing your work.)

Don’t forget to read and respond to the Honor Code instructions before you turn in your exam!

**Post your completed exam to Canvas by 9PM on Monday 26 August.**

**Suggestions for taking the exam:**

These questions are “fresh baked” for this year’s class, so there is the very real possibility that parts of them are half-baked. Contact the professor or TAs (via First Class) if something doesn’t seem right. If we do make changes and/or clarifications, we will post them in our course conference in a timely manner. PLEASE — IT IS YOUR RESPONSIBILITY TO CHECK Canvas REGULARLY!

Your best opportunity for clarification of the questions is during the exam, not afterwards. The exam questions are not intended to be ambiguous. If there are words or phrases that you do not fully understand, please ask us about them; this is not a test about American English vocabulary. You can ask any questions you like; we just may not be able to answer some questions that are too close to actual exam content.

Read carefully, and spend some time thinking before you try to answer the questions. The questions range greatly in difficulty; we suggest reading through the entire exam before you start working, so you can gauge the difficulty of the sections and budget your time.

When making assumptions about the problems, try to use the simplest set of assumptions that is consistent with all the information in the problem. Of course, more elaborate complications arise in real life, but here you’ll benefit from keeping things simple.

Partial credit IS important for some questions, so make sure your work pages clearly show your line of thinking and the specific steps of any analysis you performed. (State your assumptions! Draw your pictures!)

Good Luck.

Remember: Working this exam requires using an Excel file, which are available on Canvas.

**Exams are due to Canvas by 9PM on Monday 26 August**

***ANSWER SHEET***

**PART A** *(40 pts)*1. Simple Statistics & explanation see exam pages

2. Comment on Statistics see exam pages

3. Histogram & explanation see exam pages

4. Comment on Histogram see exam pages

5. Regression model equation: **model <- lm(formula = s\_p ~ bath + ltsz + hssz + factor(bsemt) + factor(f\_place) + factor(age5) + factor(dr) + factor(stl10) + bdrms + FAR, data = modeldata)**

6. **Best Answer for Price of House of interest: $78,467.77**

7. Prediction interval for your estimate of Price above: We should be **95% certain** that it will fall **between $42,023.06** and **$114,912.50**.

*TOTAL: = 40 pts.*

**Part A 40 points**

The Excel spreadsheet **housedata.xls** contains data on the sales of 950 single-family homes in Springfield, MA. We wish to explain and predict the price of a single-family home (Y, in thousands of dollars) using the following predictor variables:

**Data Description**

Variable Name Description **House of interest**

s\_p Sale price in dollars **?**

inv Sale date inventory of homes on market **100**

bath Number of bathrooms **2**

ltsz Lot size in acres **.25**

hssz Sq. ft. of living area **1200**

bsemt 1 if basement, 0 otherwise **0**

a\_c 1 if central a/c, 0 otherwise **1**

f\_place 1 if fireplace, 0 otherwise  **0**

garsz\_a 1 if garage, 0 otherwise **1**

dinsp 1 if dining space, 0 otherwise **1**

dw 1 if dishwasher, 0 otherwise **1**

dr 1 if dining room, 0 otherwise **0**

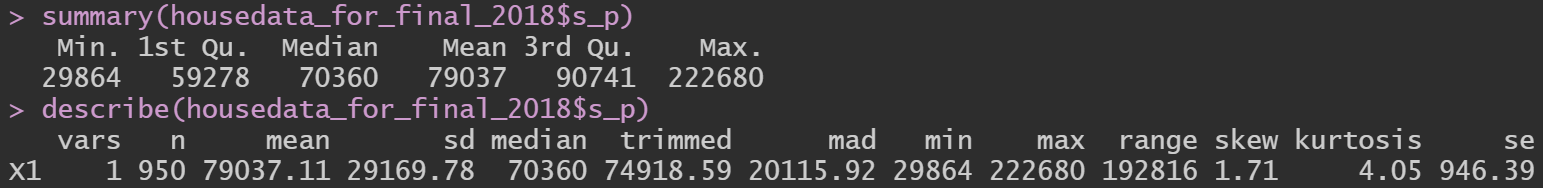
fr 1 if family room, 0 otherwise **0**

age5 1 if age <= 5 yrs, 0 otherwise **1**

stl10 1 if 1 story house, 0 otherwise **1**

bdrms Number of bedrooms **4**

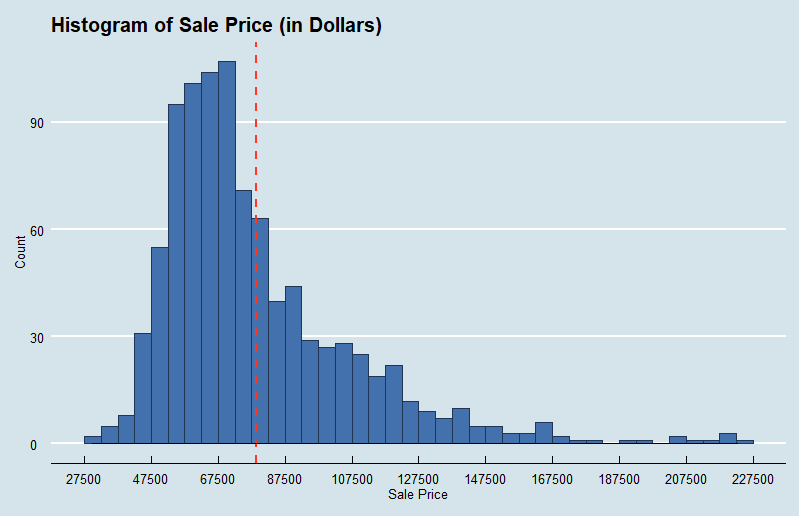
1. **Calculate simple descriptive statistics for “Sales Price”**



1. **and comment.**

Right off the bat, we can see that sales price has a huge range, with the most expensive sale price being 7.46 times more expensive than the cheapest price. With a mean of $79,037, we can also say that the dataset is skewed to the left, as midpoint between $29,864 and $22,680 is about $126,272. This is further reinforced by the low 1st quartile, median and 3rd quartile. This means that the bulk of our data will be less than ~$100,000 and that there are only a few datapoints far out close to the upper limit. (I didn’t want to call them outliers because they are still valid datapoints.) Moving down to the second set of descriptive stats, we can see that the dataset is indeed positively skewed by 1.71, meaning to the left. This also reinforces our assumption that the data has a ‘long tail’ to the right. With a kurtosis of 4.05, we can say that the data peak is quite ‘sharp’, meaning that the majority of datapoints should be closer to the mean than a normal curve. Lastly, we can look at the standard deviation and the mean absolute deviation. With the two being $29,169.78 and $20,115.92 respectively, we can say that the datapoints are pretty spread out. On average, house prices are $29,169.78 apart from each other, which is pretty high considering the average house of our dataset costs only $97,037.11.

1. **Construct a clear well labeled Histogram of “Sales Price”**



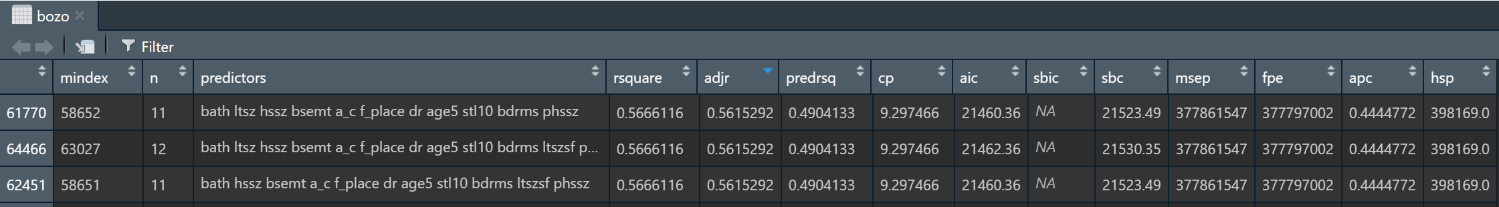
1. **and comment on what you see.**

We can see that the data is heavily skewed, with the “peak” of the curve being to the left of the dotted red line, which represents the mean. The data has a normal-ish shape but has significant outliers to the right, with the highest sale price being 7.5 times higher than the lowest sale price. Additionally, the graph shows us that the curve is ‘flatter’ than that of a regular normal bell curve, which should hint at a larger-than-normal standard deviation. We can also see that the bulk of sale prices range between $47,500 and $87,500, with approximately ¼ of the dataset sitting between $87,500 and $127,500.

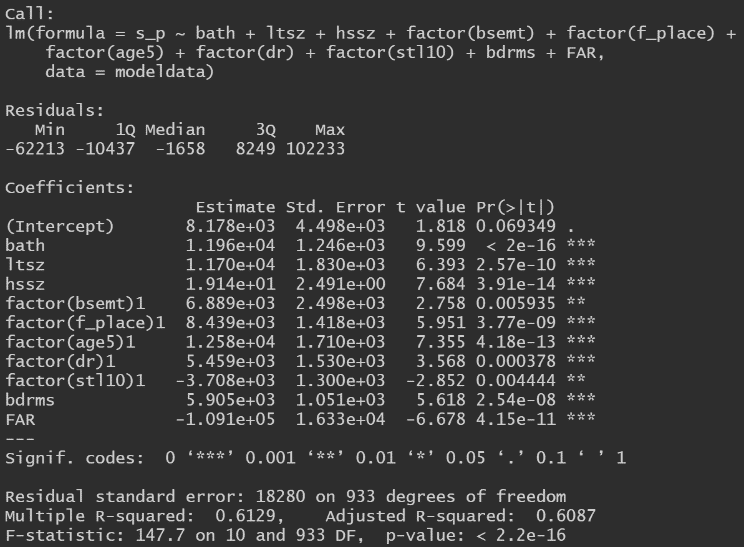
1. **Build a regression model to predict the selling price for a home. Explain your thinking and your analytical process concisely but clearly, using specific excerpts from your data analysis where appropriate Be sure to discuss any additional steps you would like to perform if you had more time for your analysis (and why those steps would be important.**

The first and most important step is to look at the data and identify any and all outliers, missing or error data. I started off by doing simple statistics analysis on all the variables given and, right off the bat, we can see that there are many unreasonable datapoints in the data. For example, one house had 20 baths while another had -4 bedrooms. By converting all lot sizes from acres to square feet, I was also able to find 5 houses that had living square feet that larger than their lot size. (Please see the large table at the end of this answer for a detailed list of comments and concerns on the variables)

After replacing all the unreasonable variables with NaN, I started creating variables that I thought would be useful for our analysis. The most notable of which is the [Floor-Area-Ratio](https://en.wikipedia.org/wiki/Floor_area_ratio), an important ratio used in house valuations. Then, I created a trial model with all the variables included and ran All-Subset-Regression, and the highest adjusted r-squared value I got was 0.5615. Using this number as a baseline for expectations but being unsatisfied with this adjusted r squared, I went on to construct my own model.



*\*A snip of the results of my ols\_step\_all\_possible call on a model that included all my variables. Bozo lives on forever!*

I started off with just lot-size and sq. ft. of living area, two variables that had good correlations with sell price, but only 0.26 correlation with each other. I then ran **resid(model)** to get the residuals of my model and appended it to the end of the dataset. By running **cor(dataset)** next, I was able to find the highest correlated variables. I then added the highest correlated variables into my model and checked to see if their **p-value** was greater than or equal to 0.05. I then removed the variables that had too high of a p-value, and/or minimally improved/negatively impacted my **adjusted R-squared**. By repeating this process, I ended up with a model that incorporated all the variables shown in the picture below. Beyond which, any additional variable had a p-value of much greater than 0.05 and/or caused my adjusted r-squared to go down.

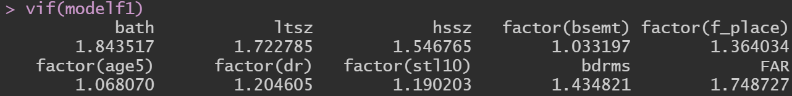
The final model has an adjusted **r-squared of 0.6087** and a **residual standard error of 18280**. The **p-value** of all our variables are **significant** as they are **below 0.05**, while **all our t-values are above ‘2’**, further reinforcing that they are significant. All the coefficients of our variables also make sense intuitively. The price of the house will be positively influenced by its lot size, sq. ft. of living area, number of baths, and number of bedrooms. Having a basement, fireplace, dining room and/or being less than 5 years old will also have a positive effect on house selling prices. Having a low floor-area-ratio and/or being single story also contributes to a higher selling price, as low FAR indicates a bigger yard, while single story houses are typically more comfortable than multiple-story houses of the same living area.

Final Model: lm(formula = s\_p ~ bath + ltsz + hssz + factor(bsemt) + factor(f\_place) + factor(age5) + factor(dr) + factor(stl10) + bdrms + FAR)

All the variables in my model are right from the dataset with exception to the floor-area-ratio, which is calculated by first converting lot size from acres to sq. ft., then dividing sq. ft. of living area by lot size (in sq. ft.). Also, changing the 0/1 variables like basement or fireplace to factors had no impact on the model as their values are only 0/1. To minimize confusion and to distinguish the 0/1 variables from the rest, I decided to keep them factored.

Coincidently, my family home is currently under renovation, and I was able to participate in all the conversations with the city hall and the architect consultant, so I gained a lot of insight into main drivers of house valuation. By far the most important factor when it comes to house prices is location. This means the neighborhood it’s in, the house’s distance to nearby schools, the ranking of the said schools, convenience, area crime rates, etc. Unfortunately, our dataset is missing this key variable. Other important factors include historical sale price of the same house, market demand (not supply), furnishing quality & materials, and age. Unfortunately, of the all highlighted above, we only have insight into house age. Even then, it’s a 1/0 variable for older than 5 years or not, which doesn’t help as the house portion of a home’s valuation linearly depreciates per year while the land holds value (excluding market influences). Bathrooms and Kitchens are the biggest drivers for furnishing costs, so assuming all the houses have comparable furnishing $/sq. ft. standards more bathrooms WILL increase house value. In reality, furnishing standards greatly differ house to house even owner to owner.

One saving grace is that we are giving the living area of the houses, which is a direct influencer of home prices. At the end of the day, we were given mostly house structure related variables while typically most of the value of a home stems from the neighborhood/location and size of its plot. A 100-year-old run-down house in a great neighborhood will easily out-value an 8-bedroom mansion in the middle of nowhere.

Running Variance Influence Factors, we can see that our values are close to 1, with a few getting close to 2. Still, this is pretty good and shows that there is no severe collinearity between our variables (something I was very worried about for the FAR variable I created).

Because of the constraint on data, there really aren’t many additional steps I can do to improve our model other than ask for more data. I spent 3 days creating transforming the existing variables and creating new dummy variables in hopes of improving the standard error, but most of them were fruitless while the few that worked resulted in variables that could not be explained. I also tried creating a neural net model using Keras and Tensorflow, but because I previously only made models for flower identification models with less than 5 discrete results, I did not have enough time to self-learn how to make a model predict a continuous result.

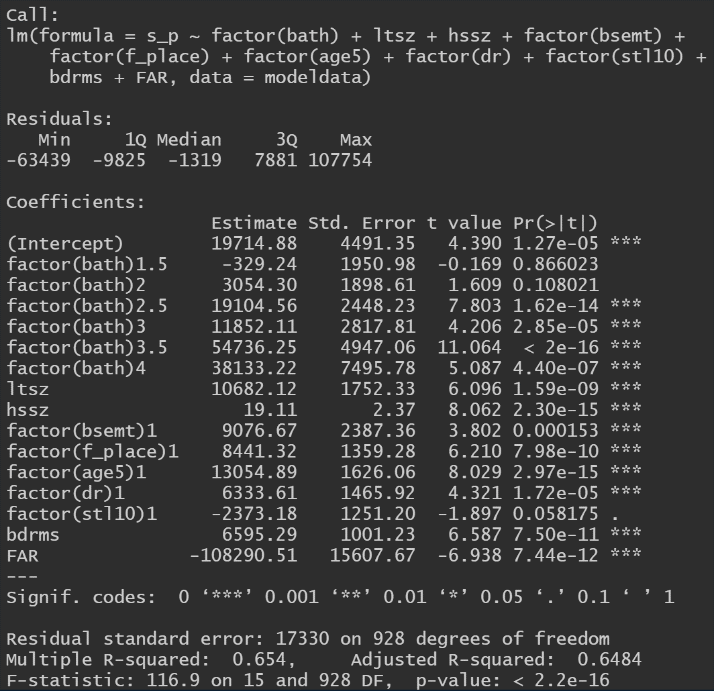
------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Additional observations, thoughts and experiments as I went through the modelling process**

------------------------------------------------------------------------------------------------------------------------------------------------------------------

The table below outlines my concerns as well as comments on the variables as I went about cleaning and analyzing them.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variable** | **Concerns** | **Comments** |
| 1. | **Sale Price ($)** | The data is heavily skewed to the left of the mean. |  |
| 2. | **Sale Date Inventory** | Pretty useless variable, as no neighborhood data is given. | Demand would be a better variable than supply. |
| 3. | **Number of Bathrooms** | One house has 20 bathrooms. | A big cost factor when constructing houses. Will be a bigger influencer of price in newer houses. |
| 4. | **Lot size in Acres** | The biggest lot size is 265x bigger than the smallest. | Land holds a lot more value than houses. Houses depreciate, land doesn’t. |
| 5. | **Sq. ft. of living area** | 5 houses have more sqft of living area than the size of acres. Of the 5, 4 are only 1 story. | Most North American zoning regulations require houses to be within ~70% of the lot size. |
| 6. | **Basement (1/0)** | The majority of houses have basements. | Not important. A terrible variable. Unfurnished basement doesn’t count as living area. Furnished does. |
| 7. | **Central A/C (1/0)** | The majority of houses have ac. | Very important in Springfield, MA, where it’s likely pretty hot in the summer and very cold in the winter. |
| 8. | **Fireplace (1/0)** |  | More of an indication of wealth than actual use for heating. Especially in houses with central A/C. |
| 9. | **Garage (1/0)** | One house has “2” for an otherwise 1/0 variable. | Size of garage would be a great variable, but alas. Also, pretty important in Springfield, as people typically park their cars indoors during the cold winters with snow and ice. |
| 10. | **Dining Space (1/0)** | Dining Space was completely missing from the data. | Is this the same variable as dining room? |
| 11. | **Dishwasher (1/0)** |  | Not a good indicator of house price. Minimal cost influence. |
| 12. | **Dining Room (1/0)** | This variable is MISSING from the dataset! | Redundant with dining space. |
| 13. | **Family Room (1/0)** |  | Pretty important, especially in western family homes. Although lack of a living room variable is disappointing. |
| 14. | **Age (1/0)** | Most houses are older than 5 years. | A very important factor that directly influences the price of the house as in the form of depreciation. Unfortunately, a 1/0 variable really doesn’t tell much. Sometimes new houses have a repair guarantee by builder, very valuable asset. |
| 15. | **House Floors (1/0)** |  | Typically, between two houses of the same living area, the 1 story house is more ‘comfortable’. This usually also indicates that the lot is bigger. |
| 16. | **Bedrooms** | 1 house has -4 bedrooms. 1 house that has 6 bedrooms but only has 1000 sqft of living space, while another house has 5 bedrooms but only with 836 sqft of living space. | An important filter people use when searching houses. More bedrooms typically mean bigger & fancier house. However, a big house in the suburbs might be the same price as a small house in downtown. A worthy variable regardless. |



Interestingly, factoring the number of bathrooms boosted the adjusted R-squared up to 0.6484 and brought down the residual standard error to $17,330, representing a +0.0397 and -950 improvement respectively. However, factoring number of bathrooms does not make a lot of sense, apart from maybe saying that houses with more bathrooms are typically larger and more luxurious, and thus should be considered differently from houses with less bathrooms which are typically smaller and cheaper. Additionally, running prediction intervals for the two models shows that the 95% prediction range only shrinks by **$3,688.97** So overall, I do not think the marginal improvement justify factoring bathrooms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Lower** | **Prediction** | **Upper** | **Range** |
| **My Model** | $42,023.06 | $78,467.77 | $114,912.50 | $72,889.44 |
| **Bathroom Factored** | $38,897.73 | $73,497.94 | $108,098.20 | $69,200.47 |

Because our knowledge of the neighborhood is severely lacking for both the house of interest and our dataset, I decided to sanity check my prediction by going onto Zillow and looking for houses with comparable characteristics to our house of interest situated in neighborhoods similar to that of Springfield, MA. I was able to find [This House](https://www.zillow.com/homedetails/2423-Dakar-Dr-Augusta-GA-30906/98529432_zpid/?) priced at $115,000 which matched every single variable of our house of interest. It is situated in [Augusta, GA](https://www.zillow.com/augusta-ga/home-values/), which is a comparable town to [Springfield, MA](https://www.zillow.com/springfield-ma/home-values/) but with a slightly lower list price/square foot. [Two houses](https://www.zillow.com/homes/for_sale/3-_beds/2.0-_baths/single_family_att/?searchQueryState=%7b%22pagination%22:%7b%7d,%22mapBounds%22:%7b%22west%22:-72.70894409516018,%22east%22:-72.39823700287502,%22south%22:42.030203872810304,%22north%22:42.17883232164953%7d,%22usersSearchTerm%22:%22springfield,%20ma%22,%22isMapVisible%22:true,%22mapZoom%22:13,%22filterState%22:%7b%22keywords%22:%7b%22value%22:%22SINGLE_FAMILY%22%7d,%22lotSize%22:%7b%22min%22:7500,%22max%22:21780%7d,%22sqft%22:%7b%22min%22:1000,%22max%22:1250%7d,%22baths%22:%7b%22min%22:2%7d,%22hasGarage%22:%7b%22value%22:true%7d,%22singleStory%22:%7b%22value%22:true%7d,%22hasAirConditioning%22:%7b%22value%22:true%7d,%22beds%22:%7b%22min%22:3%7d,%22enableSchools%22:%7b%22value%22:false%7d%7d,%22isListVisible%22:true%7d) that are situated in Springfield that closely resemble the house of interest are around $225,000 and $230,000 respectively despite only having 3 bedrooms, indicating that our model likely underestimated the price of the house in question.

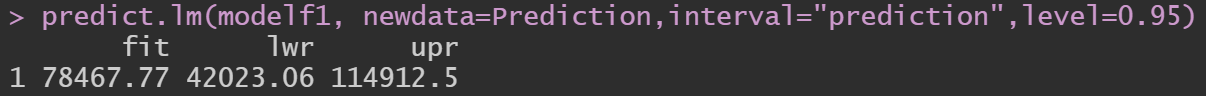
1. **What is your BEST -**MOST COMPLETE **answer to what the house of interest listed above will cost?**

A single-family home in Springfield, MA with the set of characteristics identified by “the house in question” on page 4 should on average cost around **$78,467.77**. We should be **95% certain** that it will fall **between $42,023.06 and $114,912.50.**

The price of the house will be heavily influenced by its lot size, sq. ft. of living area, number of baths, and number of bedrooms. Having a basement, fireplace, dining room and/or being less than 5 years old will have a positive effect on house selling prices. Having a low floor area ratio and/or being single story also contributes to a higher selling price.

## Prediction interval for your estimate of Price above

Running command below gives us a 95% confidence prediction interval of $42,023.06 to $114,912.50.



*\*Prediction is a dataset solely comprised of the variables for the “House of Interest”*

|  |  |  |
| --- | --- | --- |
| **Lower Bound** | **Prediction** | **Upper Bound** |
| $42,023.06 | $78,467.77 | $114,912.50 |

## END OF EXAM