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** Philipp Scherbel

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 Friday 24 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 Thursday 24 August**

***ANSWER SHEET***

**PART A** *(40 pts)*1. Simple Statistics & explanation see exam pages 5-9

2. Comment on Statistics see exam pages 9

3. Histogram & explanation see exam pages 10

4. Comment on Histogram see exam pages 10-11

5. Regression model equation see exam pages 12-20

lm(formula= logs\_p ~ltsznew +hssz + f\_place + factor(bdrms) +factor(bath) + age5 + dr + ratio+ inv)

logs\_p(predicted) = 2.812e+04 + ltsznew\*3.871e-01 + hssz\*1.397e+01 + f\_place\*9.122e+03 + factor(bdrms)2\*2.601e+03 + factor(bdrms)3\*1.008e+04 + factor(bdrms)4\*1.950e+04 + factor(bdrms)5\*2.009e+04 + factor(bdrms)6\*2.632e+04 + factor(bath)1.5\*1.811e+03 + factor(bath)2\*5.393e+03 + factor(bath)2.5\*2.176e+04 + factor(bath)3\*1.369e+04+ factor(bath)3.5\*5.886e+04 + factor(bath)4\*4.286e+04 + age5\*1.352e+04 + dr\*6.636e+03 + ratio\*-2.187e+04

6. **Best Answer for Price of House of interest**  see exam pages 21

s\_p $ 86795.44

logs\_p $ 78742.1 (with log, better r-squared)

7. Prediction interval for your estimate of Price above

fit lwr upr

s\_p $ 86795.44 $ 81659.16 $ 91931.71

logs\_p $ 78742.1 $ 74532.55 $ 83189.4 (with log, better r-squared)

With log the prediction looks more “conservative” and lower, whereas without log the skewed distribution to the right is incorporated more – and predictions higher.

*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 = 10890 sq feet**

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”

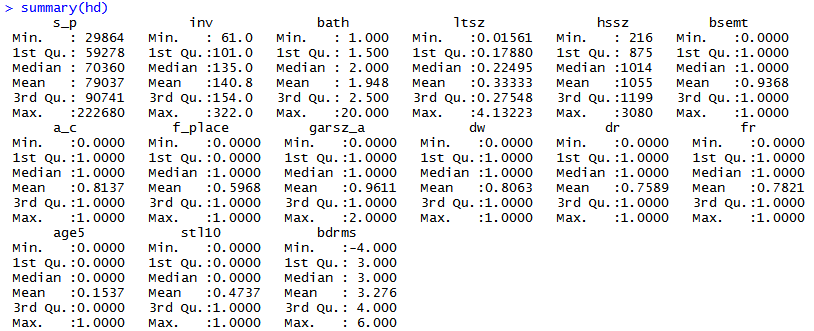
I tried to make it easy to distinguish between

* My explanations and highlights
* Copied code from r

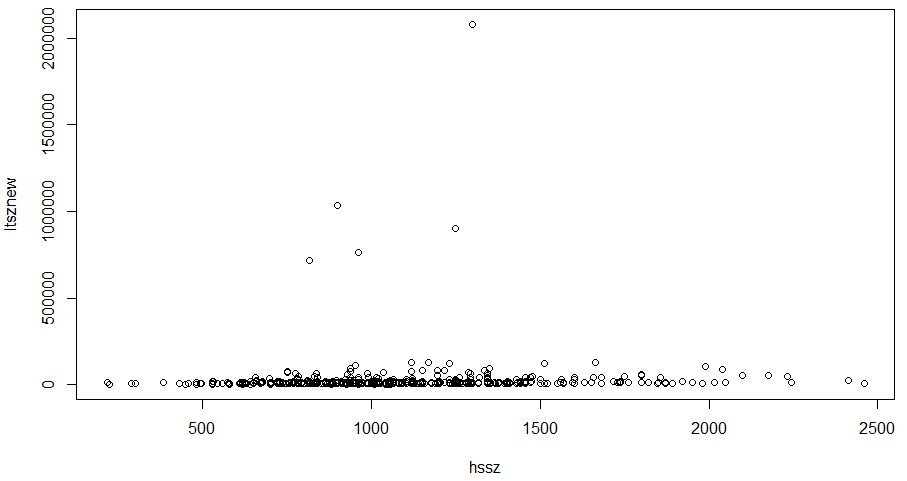
Having a look at the variables without data analysis, I would like to share my assumptions, that I will (hope to) proof right later during the actual analysis. Given demand and supply influences price, inv will influence the sale price. The less houses there are on the market, the price of the remaining ones will increase (negative correlated). In addition, the bigger the house is (given same area, age etc.) the higher the price. My guess it that the size of the living area is actually not as important as the amount of beds and bathrooms as people usually search for specific amounts of these rooms first, i.e. a house with 2 bedrooms and a huge living room most likely will sell at a lower price compared to a house with 3-4 bedrooms and a smaller living room. From the binary variables, I think that the age matters most – and the more expensive the houses are, also a fireplace may be important to the potential buyers of more expensive houses.

To start off, I will run:

* Describe() and summary()
* Display scatter plots
* Check correlations
* Check outliers/ faulty data



The summary shows that many variables have a very high max, which implies one specific large house (lot size and house size) or it could be actually false data. As the units are different, I converted also the lot sizes (in acre) to sq feet (ltsznew). Another record shows 20 baths. In addition, a negative amount of bedrooms is not possible, thus we have wrong data here. I will investigate the house and lot size via plotting and replace the data-point with bath = 20 to the median (I assume that way the wrong number will still be in the dataset calculating the median but it should be fine – compared to using the mean). Also, garage has max of 2 but it is a binary variable with 0 or 1 only, so I will set it to median of that column.



As the graph shows, there are several house/lot combinations far from the “bulk”, thus I decided to delete hssz > 2200 and ltsznew > 500000 (acre converted to sqfeet) and to replace each with the column’s median.

hd$ltsznew <- hd$ltsz\*43560

plot(hd$ltsznew, hd$hssz)

#I decided some are outliers or just wrong data and to replace them with median

hd$hssz[hd$hssz > 2200 ] <- median(hd$hssz)

hd$ltsznew[hd$ltsznew > 500000 ] <- median(hd$ltsznew)

hd$garsz\_a[hd$garsz\_a > 1 ] <- median(hd$garsz\_a)

hd$bdrms[hd$bdrms < 0 ] <- median(hd$bdrms)

hd$bath[hd$bath > 10 ] <- median(hd$bath)

plot(hd$ltsznew, hd$hssz)

I looked at the ratio from hssz to ltsznew and 5 houses were bigger than the lot. I had two options, deleting them or changing. For deleting I used the following formula:

hd$ratio <- hd$hssz/hd$ltsznew

summary(hd$ratio)

hd <- hd[hd$ratio<1, ]

#test if it worked

hd$ratio[hd$ratio>1]

shows numeric 0

but since the dataset is relatively small, I looked into these 5 ratios and decided to multiply the lotsizes with 10, as this would match with general median and ratios.

hd$ratio <- hd$hssz/hd$ltsznew

hd$ltsznew[hd$ratio >1] <- hd$ltsznew[hd$ratio>1]\*10

hd$ratio[hd$ratio > 1]

[1] 1.304348 1.142857 1.547619 1.200000 1.273469

hd$ltsznew[hd$ratio>1]

[1] 690 840 840 680 980

hd$hssz[hd$ratio>1]

[1] 900 960 1300 816 1248

mean(hd$hssz)

[1] 1051.398

mean(hd$ltsznew)

[1] 13873.53

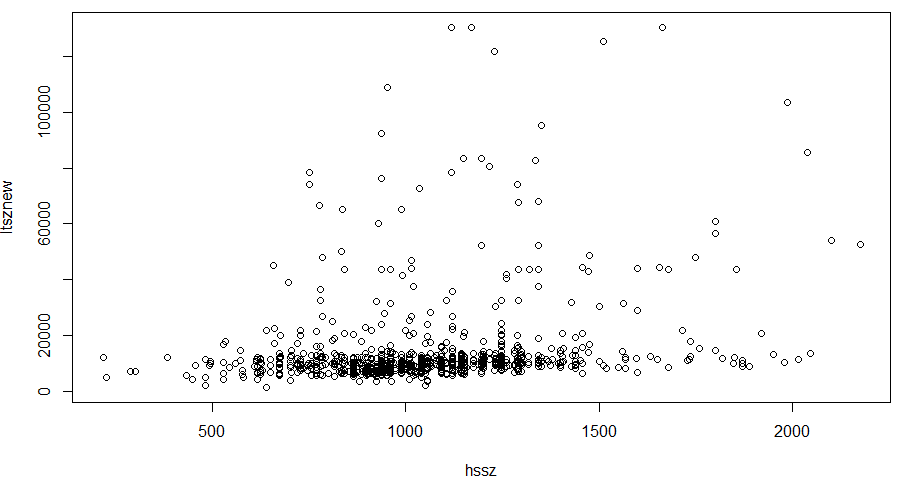
median(hd$hssz)

[1] 1014

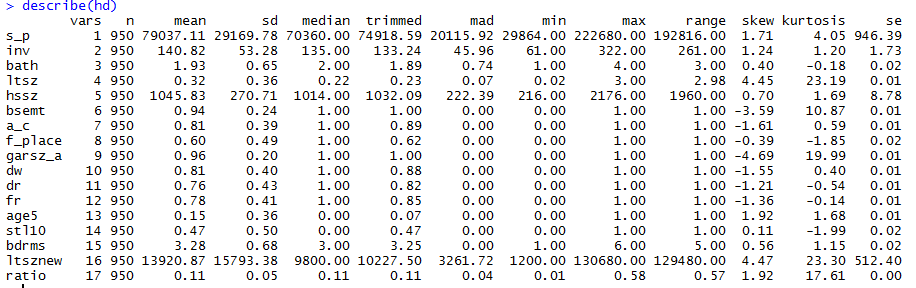
median(hd$ltsznew)

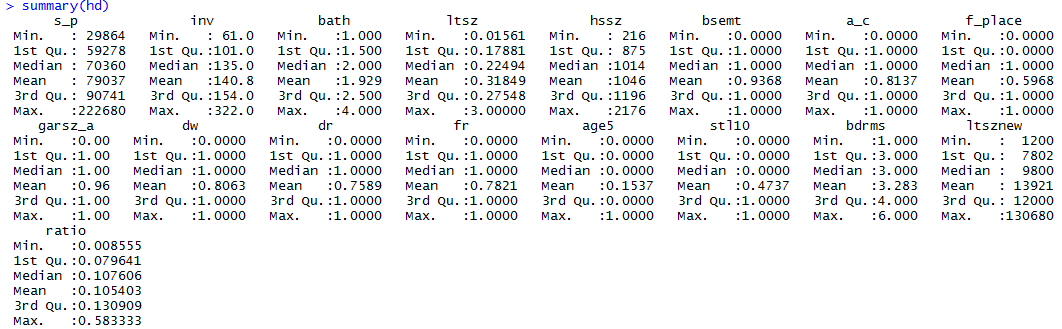
[1] 9798.5

This was the outcome and I decided to not “clean” further.

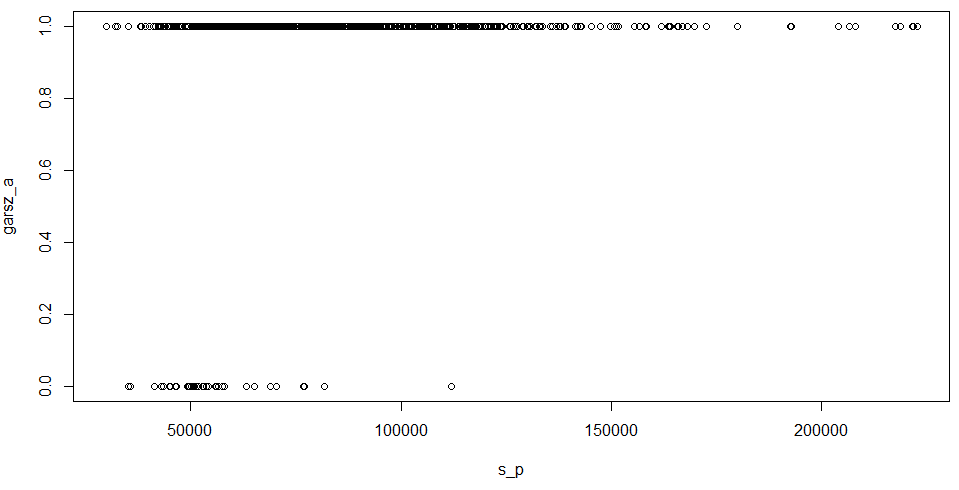


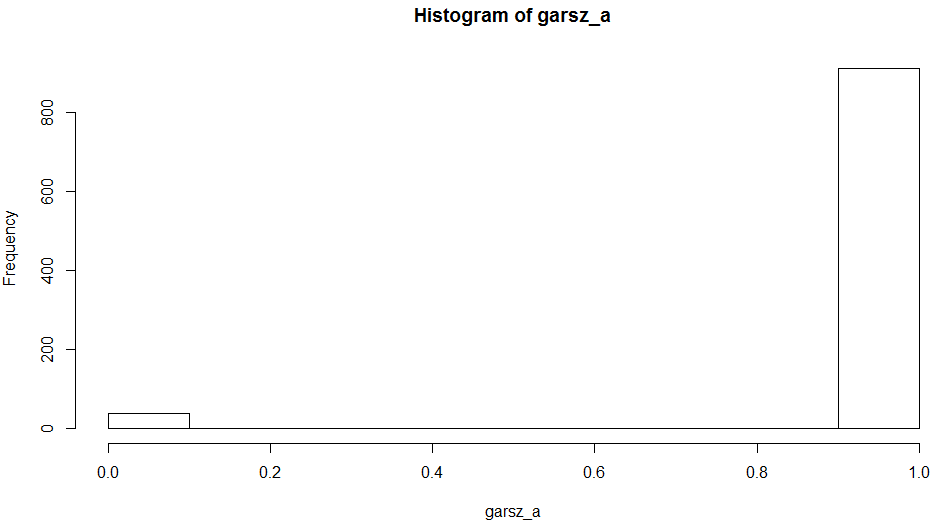
after cleaning, describe() and summary () look as follows:





Looking at the max, there are still huge lots, but for an amateur real estate agent like me, the data looks fine. There are no negative values or impossible values (2 for binary for example) any more, and mean and median are closer now for each variable which means less outliers. Some values are heavily skewed like garage and basement. But this is due to their binary character.





1. and comment.

As I needed to perform cleaning before presenting describe and summary, I commented throughout 1)

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

# histogram

hist(s\_p)

par(mar = c(5, 4, 4, 5))

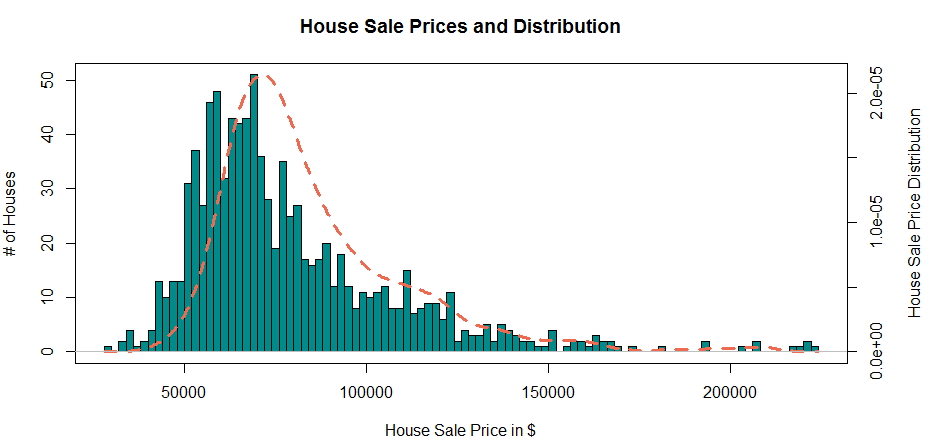
hist(hd$s\_p, nclass=100, type ="l", ylab = "# of Houses", main = "House Sale Prices and Distribution", xlab = "House Sale Price in $", col = "cyan4")

par(new = TRUE)

plot(density(hd$s\_p), type = "l", xaxt = "n", yaxt = "n", ylab = "", main="", xlab = "", col = "coral2", lty = 2, lwd = 3)

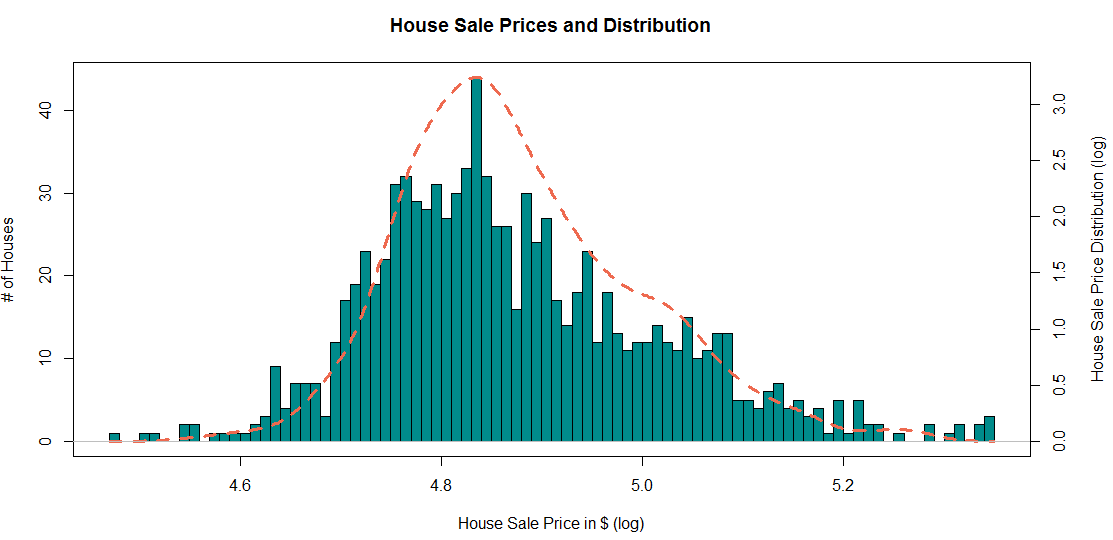
axis(side = 4)

mtext("House Sale Price Distribution", side = 4, line = 3)



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

The histogram shows that many houses cost around 50-75 k (“the most frequent house”, i.e. median house sold, costs around 70k) and then there are more expensive houses, which are less frequent, the more expensive they are. These houses move the mean to the right, as there are not many houses cheaper than 50 k who would pull it to the left. Overall the curve looks somehow normal from 0 to 100 k but then extends with more expensive houses to the far right on the x axis (skew). This is a reason to use log and I calculated log of s\_p to see if the distribution looks more normal.

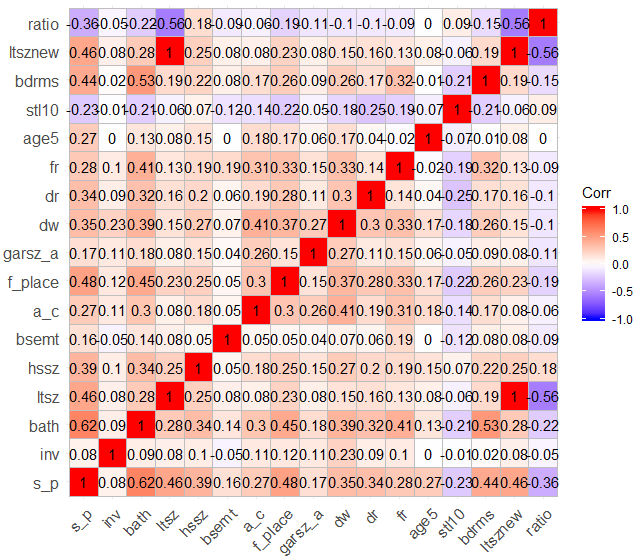


It helps making the distribution look more normal but I don’t expect the impact on the model be very high as the distribution still looks very scattered/ random and rather steep left compared to the right side.

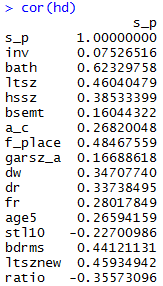
1. Build a regression model to predict the selling price for a home. Explain your thinking and your analytical process concisely but clearly, using specific excerpts from your data analysis where appropriate Be sure to discuss any additional steps you would like to perform if you had more time for your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

I created a heatmap and the correlation table first to decide with which variables I want to move forward.

ggcorrplot(corr, hc.order = TRUE, type = "lower", lab = TRUE)



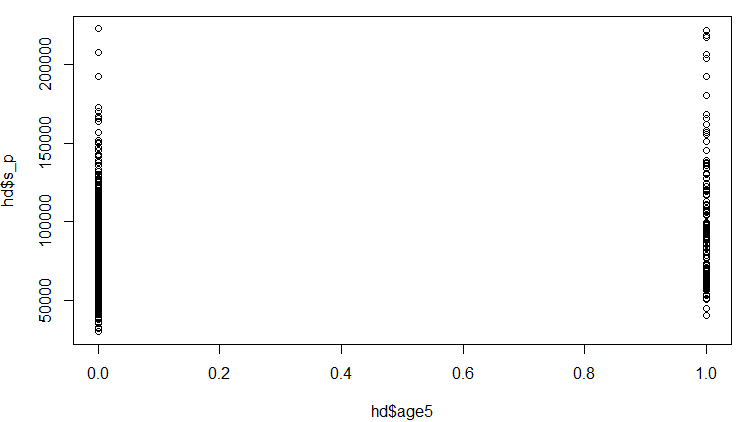
cor(hd)



These variables are highly correlated with s\_p:

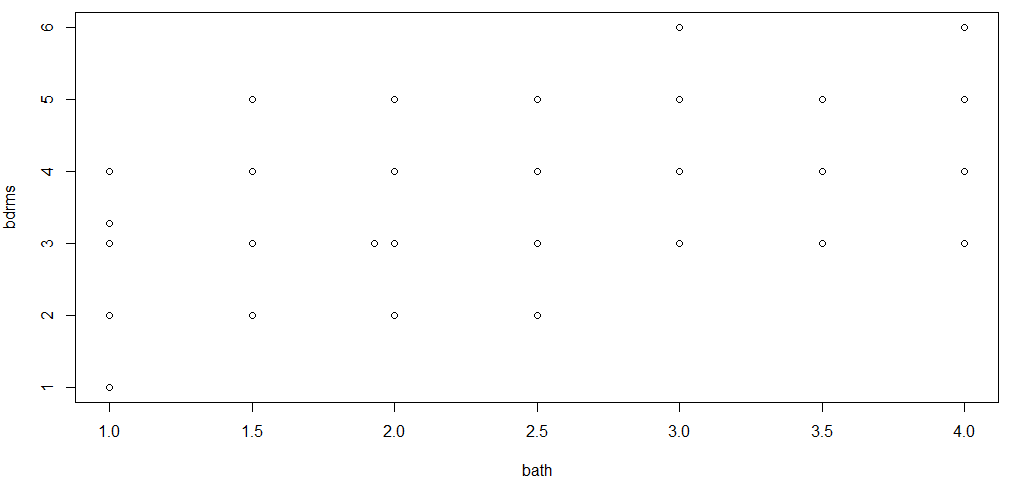
* Ltsz(new)
* Hssz
* Bath
* Bdrms
* F\_place
* Ratio

In addition, age5 should be very important but it is not highly correlated. Plotting s\_p and age5 doesn’t show a significant connection.



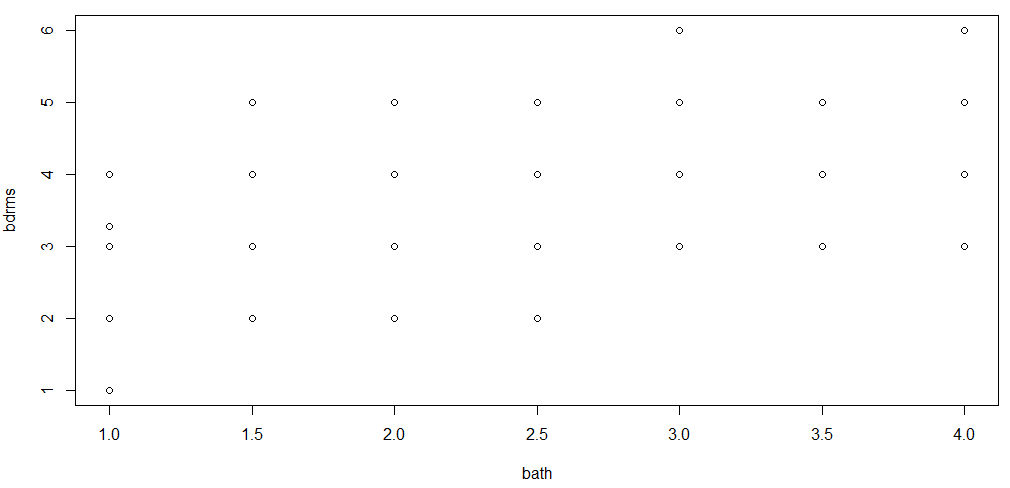
To see if bdrms - bath are related and ltsznew - hssz, I will plot both. As bdrms and bath increase linear with a pretty constant ratio, I will only include the variable higher correlated with s\_p. Ltsznew and hssz, on the other hand, do not seem highly correlated. As Bath and Bedroom may be factored variables and “jump” without clear linear connection (there won’t be for example 2,43 baths possible), I will add bath as factor the model instead of linear later. I tried factor(bath) to see what these factors are and not all values are ending with .0 or .5 so I set the one outlier (1.93..) to 2.

Before:



After:

hd[hd$bath>1.9 & hd$bath <2] <- 2



As this was just a single value, I didn’t run describe and summary again and also the correlation table shouldn’t change much. The cleaning helped rather to not get additional factors moving forward:

To start off, I tried both bath and factor(bath) as bath was correlated the strongest with s\_p.

My first model:

Call:

model1 <- lm(s\_p~bath)

hd$resid1 <- resid(model1)

summary(model1)

Call:

lm(formula = s\_p ~ bath)

Residuals:

Min 1Q Median 3Q

-65290 -13987 -3519 8890

Max

122462

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24985 2324 10.75 <2e-16\*\*\*

bath 28014 1141 24.54 <2e-16\*\*\*

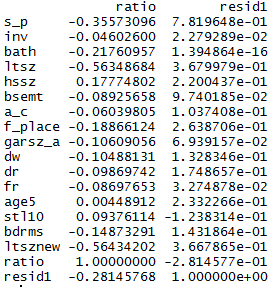
Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22820 on 948 degrees of freedom

Multiple R-squared: 0.3885, Adjusted R-squared: 0.3879

F-statistic: 602.4 on 1 and 948 DF, p-value: < 2.2e-16



And with factor (bath):

summary(model1)

Call:

lm(formula = s\_p ~ factor(bath))

Residuals:

Min 1Q Median 3Q Max

-64662 -12735 -3403 8293 121339

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 55882 1692 33.031 < 2e-16 \*\*\*

factor(bath)1.5 10561 2317 4.558 5.84e-06 \*\*\*

factor(bath)2 21401 2077 10.303 < 2e-16 \*\*\*

factor(bath)2.5 48040 2508 19.151 < 2e-16 \*\*\*

factor(bath)3 44270 2850 15.536 < 2e-16 \*\*\*

factor(bath)3.5 95593 5753 16.617 < 2e-16 \*\*\*

factor(bath)4 76092 9137 8.328 2.87e-16 \*\*\*

---

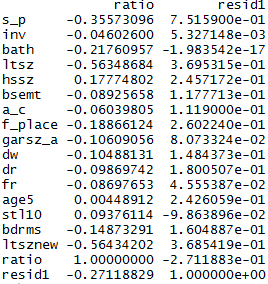
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21990 on 943 degrees of freedom

Multiple R-squared: 0.4351, Adjusted R-squared: 0.4315

F-statistic: 121.1 on 6 and 943 DF, p-value: < 2.2e-16

The standard error is already lower and I will look at the residuals:



There is no clear “winner” here for the next variable to put in the model. Also the standard error is very high, so testing residuals and modeling and repeating may not be the best way here. Thus, I chose several variables that are the most correlated here and added them.

* Ratio
* Ltsznew
* Hssz
* Age5
* F\_place

Model2:

Call:

lm(formula = s\_p ~ ltsznew + hssz + f\_place + factor(bath) + age5 + ratio)

Residuals:

Min 1Q Median 3Q Max

-62835 -10397 -1545 7705 107509

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.476e+04 2.858e+03 15.662 < 2e-16 \*\*\*

ltsznew 2.559e-01 5.087e-02 5.031 5.86e-07 \*\*\*

hssz 2.189e+01 2.686e+00 8.150 1.16e-15 \*\*\*

f\_place 1.008e+04 1.405e+03 7.176 1.46e-12 \*\*\*

factor(bath)1.5 4.534e+03 1.971e+03 2.301 0.0216 \*

factor(bath)2 9.485e+03 1.863e+03 5.090 4.33e-07 \*\*\*

factor(bath)2.5 2.783e+04 2.376e+03 11.712 < 2e-16 \*\*\*

factor(bath)3 2.306e+04 2.662e+03 8.662 < 2e-16 \*\*\*

factor(bath)3.5 7.122e+04 4.982e+03 14.298 < 2e-16 \*\*\*

factor(bath)4 5.287e+04 7.719e+03 6.848 1.35e-11 \*\*\*

age5 1.215e+04 1.690e+03 7.186 1.36e-12 \*\*\*

ratio -1.158e+05 1.722e+04 -6.728 2.99e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18330 on 938 degrees of freedom

Multiple R-squared: 0.6099, Adjusted R-squared: 0.6053

F-statistic: 133.3 on 11 and 938 DF, p-value: < 2.2e-16

After, I tried different models and added binary variables and found this combination:

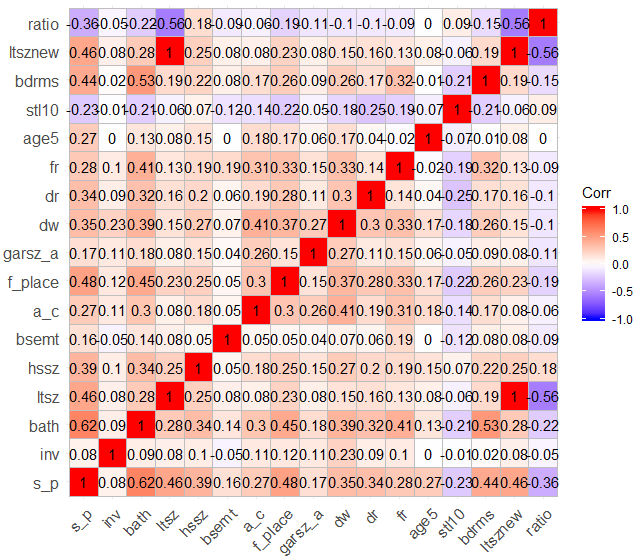
Model3:

Call:

lm(formula = s\_p ~ ltsznew + hssz + f\_place + factor(bath) + age5 + dr + ratio)

with RSE: 18140

As this path with residuals doesn’t seem successful, I checked the original heatmap again and added factor(bdrms).



Model4:

Call:

lm(formula = s\_p ~ ltsznew + hssz + f\_place + factor(bdrms) + factor(bath) + age5 + dr + ratio)

Residuals:

Min 1Q Median 3Q Max

-63864 -10117 -1076 7656 103120

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.984e+04 9.310e+03 4.279 2.07e-05 \*\*\*

ltsznew 2.411e-01 4.944e-02 4.877 1.27e-06 \*\*\*

hssz 1.990e+01 2.629e+00 7.569 9.01e-14 \*\*\*

f\_place 8.785e+03 1.381e+03 6.361 3.15e-10 \*\*\*

factor(bdrms)2 -2.088e+03 9.226e+03 -0.226 0.821

factor(bdrms)3 4.951e+03 9.049e+03 0.547 0.584

factor(bdrms)3.28345626975764 2.218e+03 1.992e+04 0.111 0.911

factor(bdrms)4 1.414e+04 9.164e+03 1.543 0.123

factor(bdrms)5 1.429e+04 9.501e+03 1.504 0.133

factor(bdrms)6 1.914e+04 1.605e+04 1.193 0.233

factor(bath)1.5 8.675e+02 2.004e+03 0.433 0.665

factor(bath)2 4.056e+03 1.952e+03 2.078 0.038 \*

factor(bath)2.5 1.962e+04 2.512e+03 7.811 1.53e-14 \*\*\*

factor(bath)3 1.330e+04 2.862e+03 4.646 3.86e-06 \*\*\*

factor(bath)3.5 6.034e+04 5.001e+03 12.066 < 2e-16 \*\*\*

factor(bath)4 4.082e+04 7.907e+03 5.163 2.97e-07 \*\*\*

age5 1.377e+04 1.657e+03 8.308 3.41e-16 \*\*\*

dr 6.794e+03 1.473e+03 4.613 4.52e-06 \*\*\*

ratio -1.124e+05 1.676e+04 -6.704 3.51e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17730 on 931 degrees of freedom

Multiple R-squared: 0.6376, Adjusted R-squared: 0.6305

F-statistic: 90.98 on 18 and 931 DF, p-value: < 2.2e-16

Bedrooms also has values that don’t match integers, thus I need to clean the data a little further.

Factor(bdrms)

hd$bdrms[8] <- 3

As bath 1.5 is not very high correlated, I will try two things and add new columns:

* Round bath 1.5 down to 1.5
* Round bath 1.5 up to 2

The standard error went up to 18550. The model got simpler, but slightly worse, so I will keep the step 1.5 for bath. At this point I don’t know if getting rid of additional variables is possible. But f\_place is not very helpful in terms of decreasing the error and models should be kept simple – if possible.

From here, I tried to add (or get rid of) binary variables to improve (or at least simplify the model):

Adding inv to the model:

lm(formula = s\_p ~ ltsznew + hssz + f\_place + factor(bdrms) + factor(bath) + age5 + dr + ratio + inv)

RSE: 17710 instead of 17730

I will keep inv in there.

So my final model looks like this:

Call:

lm(formula = s\_p ~ ltsznew + hssz + f\_place + factor(bdrms) + factor(bath) + age5 + dr + ratio + inv)

Residuals:

Min 1Q Median 3Q Max

-61580 -10007 -1204 7864 109318

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.041e+04 9.166e+03 3.318 0.000943 \*\*\*

ltsznew 3.908e-01 3.453e-02 11.317 < 2e-16 \*\*\*

hssz 1.421e+01 2.378e+00 5.976 3.25e-09 \*\*\*

f\_place 9.258e+03 1.380e+03 6.708 3.42e-11 \*\*\*

factor(bdrms)2 2.547e+03 9.181e+03 0.277 0.781548

factor(bdrms)3 1.014e+04 8.995e+03 1.128 0.259792

factor(bdrms)4 1.950e+04 9.108e+03 2.141 0.032509 \*

factor(bdrms)5 1.967e+04 9.457e+03 2.080 0.037794 \*

factor(bdrms)6 2.544e+04 1.600e+04 1.590 0.112133

factor(bath)1.5 2.356e+03 2.016e+03 1.169 0.242859

factor(bath)2 5.459e+03 1.932e+03 2.826 0.004810 \*\*

factor(bath)2.5 2.211e+04 2.489e+03 8.885 < 2e-16 \*\*\*

factor(bath)3 1.396e+04 2.856e+03 4.889 1.19e-06 \*\*\*

factor(bath)3.5 5.951e+04 5.013e+03 11.871 < 2e-16 \*\*\*

factor(bath)4 4.301e+04 7.887e+03 5.453 6.36e-08 \*\*\*

age5 1.348e+04 1.649e+03 8.172 9.83e-16 \*\*\*

dr 6.718e+03 1.471e+03 4.566 5.63e-06 \*\*\*

ratio -2.200e+04 6.230e+03 -3.531 0.000434 \*\*\*

inv -2.104e+01 1.118e+01 -1.881 0.060276 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17710 on 931 degrees of freedom

Multiple R-squared: 0.6384, Adjusted R-squared: 0.6314

F-statistic: 91.3 on 18 and 931 DF, p-value: < 2.2e-16

In addition, going back to the histogram and log, I used logs\_p and the result was:

Call:

lm(formula = logs\_p ~ ltsznew + hssz + f\_place + factor(bdrms) + factor(bath) + age5 + dr + ratio + inv)

Residuals:

Min 1Q Median 3Q Max

-0.36501 -0.05177 -0.00419 0.05274 0.33324

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.555e+00 4.258e-02 106.962 < 2e-16 \*\*\*

ltsznew 1.567e-06 1.604e-07 9.766 < 2e-16 \*\*\*

hssz 7.184e-05 1.105e-05 6.502 1.29e-10 \*\*\*

f\_place 5.827e-02 6.411e-03 9.089 < 2e-16 \*\*\*

factor(bdrms)2 6.250e-02 4.265e-02 1.465 0.143196

factor(bdrms)3 1.153e-01 4.179e-02 2.758 0.005928 \*\*

factor(bdrms)4 1.565e-01 4.231e-02 3.698 0.000230 \*\*\*

factor(bdrms)5 1.663e-01 4.393e-02 3.786 0.000163 \*\*\*

factor(bdrms)6 2.216e-01 7.433e-02 2.981 0.002947 \*\*

factor(bath)1.5 2.567e-02 9.366e-03 2.741 0.006244 \*\*

factor(bath)2 4.960e-02 8.973e-03 5.527 4.22e-08 \*\*\*

factor(bath)2.5 1.205e-01 1.156e-02 10.419 < 2e-16 \*\*\*

factor(bath)3 8.829e-02 1.327e-02 6.655 4.84e-11 \*\*\*

factor(bath)3.5 2.457e-01 2.329e-02 10.550 < 2e-16 \*\*\*

factor(bath)4 1.835e-01 3.664e-02 5.009 6.55e-07 \*\*\*

age5 5.354e-02 7.662e-03 6.988 5.30e-12 \*\*\*

dr 4.231e-02 6.835e-03 6.190 9.00e-10 \*\*\*

ratio -1.088e-01 2.894e-02 -3.758 0.000182 \*\*\*

inv -1.496e-04 5.196e-05 -2.880 0.004070 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08228 on 931 degrees of freedom

Multiple R-squared: 0.6668, Adjusted R-squared: 0.6604

F-statistic: 103.5 on 18 and 931 DF, p-value: < 2.2e-16

As the r-squared is higher in this model, I will use this model as final!

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

Final model:

lm(formula= s\_p ~ltsznew+hssz+f\_place + factor(bdrms)+factor(bath)+age5+dr+ratio+inv)

Recalling 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 = 10890 sq feet**

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**

additional variable:

ratio: 1200/20890 = 0.05744375

Calculating:

Without log:

predict.lm(model5, newdata=data.frame(inv=100,bath=2,ltsznew=10890,f\_place= 0, hssz=1200, age5=1, ratio= 0.05744375, bdrms=4, dr =0), interval="confidence", level=0.95)

fit lwr upr

1 86795.44 81659.16 91931.71

With log (better r-squared)

predict.lm(model7, newdata=data.frame(inv=100,bath=2,ltsznew=10890,f\_place= 0, hssz=1200, age5=1, ratio= 0.05744375, bdrms=4, dr =0), interval="confidence", level=0.95)

fit lwr upr

1 4.896207 4.872346 4.920068

There were 50 or more warnings (use warnings() to see the first 50)

fit <- 10\*\*4.896207

fit

[1] 78742.1

lwr <- 10\*\*4.872346

lwr

[1] 74532.55

upr <- 10\*\*4.920068

upr

[1] 83189.

## END OF EXAM