1. What is the response variable in this analysis? Briefly explain. [2pts]
   - The response variable in the dataset is the price of a home. This is because we are looking at how the other variables affect the price of a home.
2. Check variable type for each variable and change type if needed. [1pt]
   - Price: numerical
   - Bed: numerical
   - Bath: numerical
   - Area: numerical
   - Year built: numerical
   - Cooling: categorical
   - Lot: numerical
3. Provide the population regression model/line for this analysis making sure to use correct notation. Information on inserting Greek letters in Google docs is found here. It is expected you will use Greek letters and subscripts correctly (e.g. beta_0 is NOT acceptable!) [2pts]

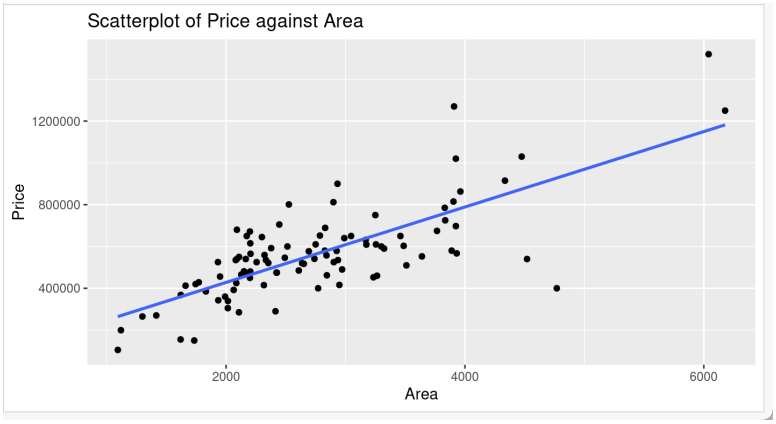
Price = $\beta_0 + \beta_{bed} + \beta_{bath} + \beta_{area} + \beta_{year\_built} + \beta_{cooling} + \beta_{lot} + \varepsilon$

4. Let's first look at the relationship between sale price and area of the home.

   a. Construct and interpret the appropriate plot to examine the relationship between these two variables. Make sure to appropriately place the variables in this plot! [5pts]

   b. Based on this plot, explain if you think the area is useful in predicting sale price. [2pts]

   The relationship between area and price is positive, linear, and moderately strong. There are a couple potential outliers including a house with an area of about 5800 square feet with a price of $400,000 and another house with an area of about 6100 square feet with a price of about $15,000,000. Based on this plot, we can conclude that area does have an effect on the price of a house.

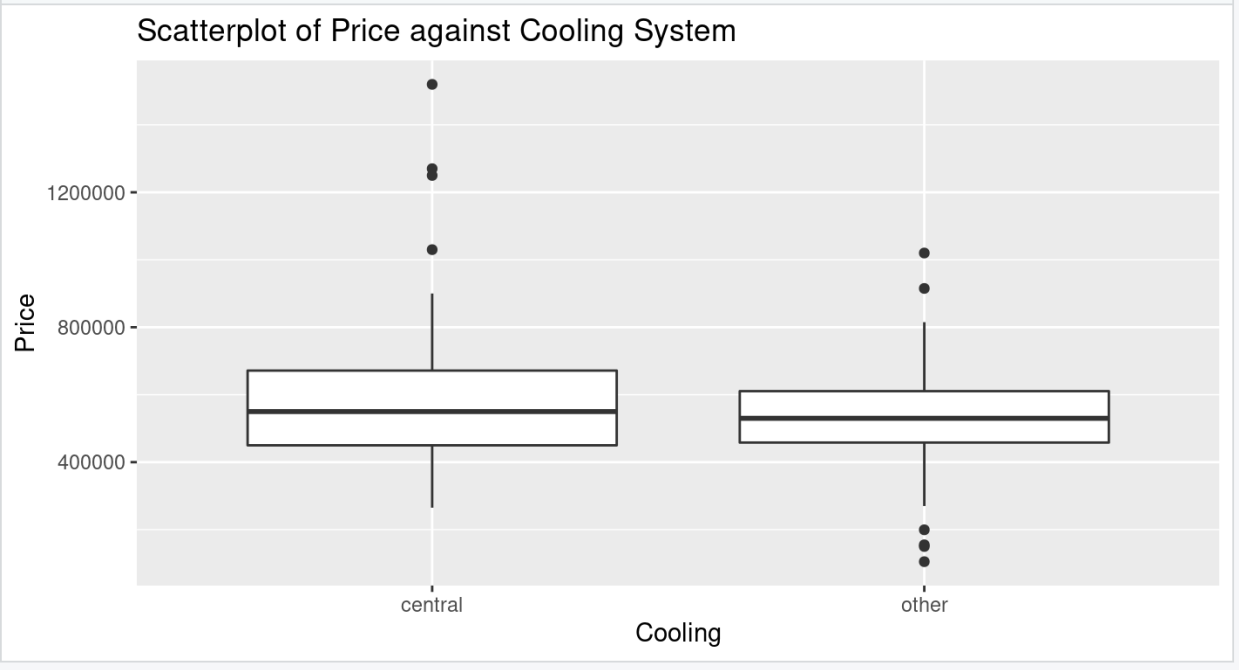


5. Next let’s look at the relationship between price and type of cooling system.

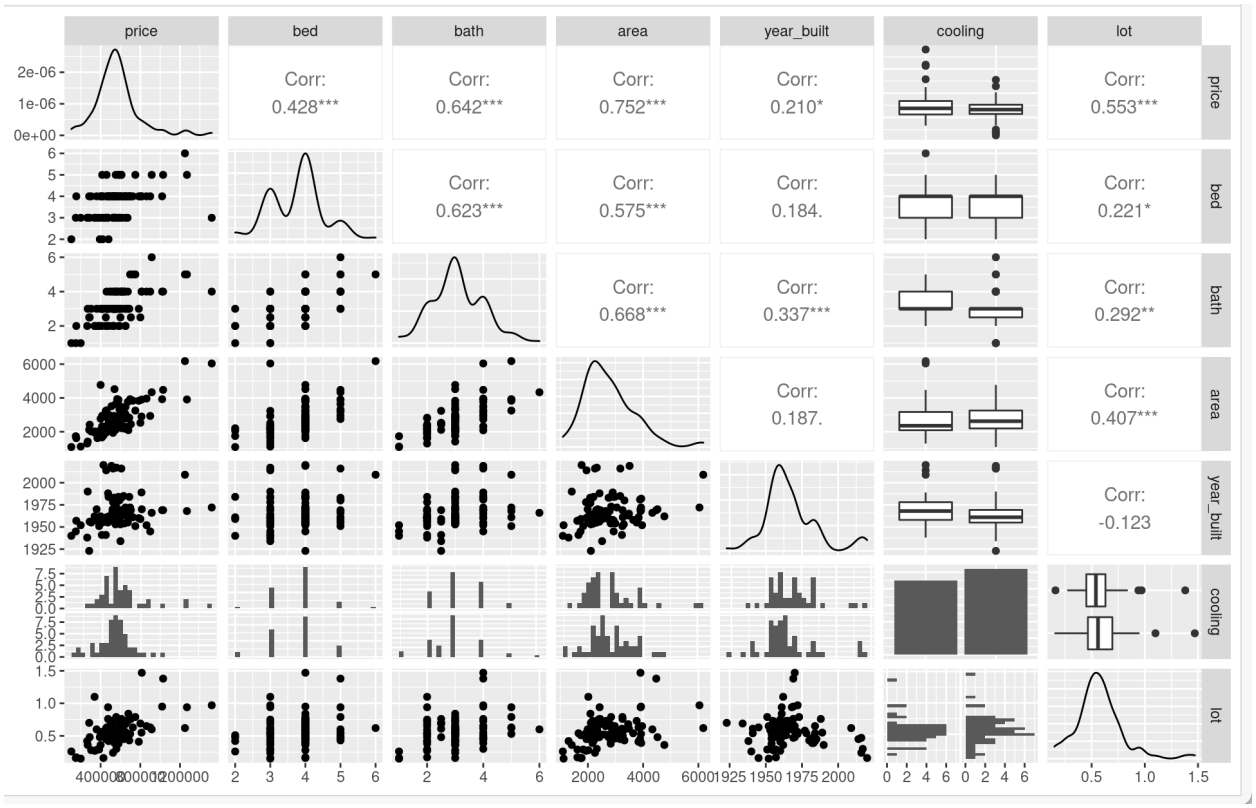
a. Construct and interpret the appropriate plot to examine the relationship between these two variables. Make sure to appropriately place the variables in this plot![5pts]

b. Based on this plot, explain if you think type of cooling system is useful in predicting sale price.[2pts]

When comparing the distribution of price of houses by cooling system, the most noticeable difference is the mean price for houses with a central cooling system is higher than houses with another type of cooling system ($610687 compared to $524886). The shape of the distribution is very similar for both, as they are both skewed left. The houses with another type of cooling systems have a slightly smaller spread than those with a central cooling system. There are a couple outliers for both cooling groups. Cooling systems don't have a huge effect on the sales price of a home because it ultimately depends on the person's preference.



6. Obtain the scatterplot matrix for this dataset. Make sure to create the plot so that the last row of the scatterplot matrix has price on the y-axis and the other variables on the x-axis. [1pt]



a. What (numerical) variable is most highly correlated with the response? Provide both the variable name and the correlation.[1pt]

- Area is most highly correlated with the response variable (0.752)

b. What pair of (numerical) variables is least correlated? Provide variable names and the correlation. [1pt]

- Year built is least correlated with the response variable (0.210)

c. Why are no values for correlation provided when considering the variable cooling? [2pts]

- No values are provided for correlation with cooling because this is a categorical variable.

7. Now that we've looked at the data, let's fit our regression.

a. Obtain the parameter estimates for our linear regression model. Report the estimated regression line—make sure to use correct notation, variable names, and convert numbers out of RStudio's scientific notation! Information on how to add a ^ over a letter or word in Google docs is found here. [2pts]

$\widehat{price}$ =
-2539836-31568.8bed+50852.8bath+127.1area+1308.8year_built-68476.3cooling+321592.5lot

For each additional **square foot** in area of the home, it is **estimated** that, on **average**, sales price **increased** by **$127.1** holding **bed,bath,year built,cooling and lot** constant.

On **average**, it is **estimated** that houses with a central cooling system sell for $**68476.3 less** than houses with a different type of cooling system holding **bed,bath,area,year built,and lot** constant.

8. We expect that homes that are larger in area would have a higher selling price. Conduct the appropriate hypothesis test to test this by completing the following parts. a. Provide the null and alternative hypotheses of interest. [1pt]

H0:βarea=0

HA:βarea>0

b. Provide the correct test statistic and associated df. [0.5pt]

t=6.418 df=89

c. Provide the associated p-value. [0.5pt]

pvalue=<0.0001

d. Provide your decision making sure to justify your decision. [1pt]

Reject the null hypothesis because 0.05<0.0001

e. Provide your conclusion. [2pts]

At the 5% level of significance we reject the null hypothesis and conclude that there is very strong evidence that area has a positive effect on price after accounting for bed, bath, year built, cooling, and lot. (t=6.418, df=89, p-value<0.0001)

)

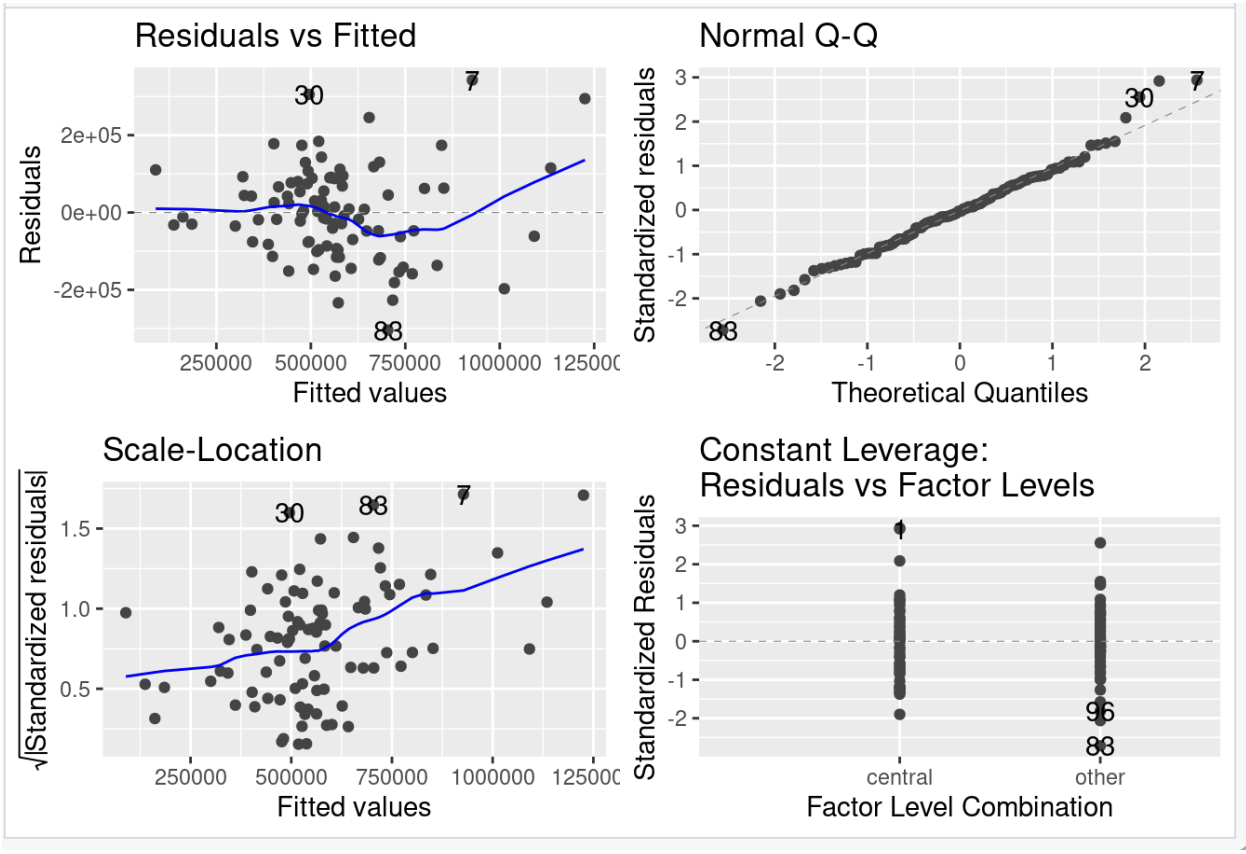
9. We are also interested in the impact that cooling system has on sales price. Report and interpret the 95% confidence interval associated with the cooling system. [3pts]

We are 95% confident that on average, the average sales price of a home for houses with a cooling system are $119,271 to $17,681. lower than the average sales price for a home with another type of cooling system after controlling for bed, bad, area, year built, and lot.

10. Silvio and his family sold their four bedroom, 3 bath house 2624 square foot home in November 2020. The home was built in 1992 with a central cooling system and a 0.35 acre lot. How much would you predict Silvio and his family sold their house for? [2pts]

- $535,650.4

11. For our results to be valid, there are four assumptions that must be satisfied.



Linearity: the relationship between __predictor_____ and ___response_ is __ linear___. Based on __the residual plot___ this assumption __is not__ reasonably satisfied because _the blue smoother line is not flat at y=0__.

Independence: the ___observations____ are ___independent__ of one another. Based on __the backround__ this assumption ___is not___ reasonably satisfied because there is no random sampling, data were not collected over time, but were collected over space and in groups/clusters___.

a. iii. Normality: __The response__ follows a _normal__ distribution. Based on ___the QQplot___ this assumption _is_ reasonably satisfied because __the points are close to the line____.
i. iv. Equal variance: for each value (or set of values) for the __explanatory__ variable(s), the variability of ___the response___ is ___the same__. Based on ___the residual plot_____ this assumption __is not_ reasonably satisfied because _there is a fan shape in the distribution__.