Marks

Student ID:

| | | 2 | 6 | 1 | 0 | 3 | 6 | 8 |
|---|---|---|---|---|---|---|---|---|

Name Pediaditis Georgios

Final Model Chosen Price=58402 + 175635*bath

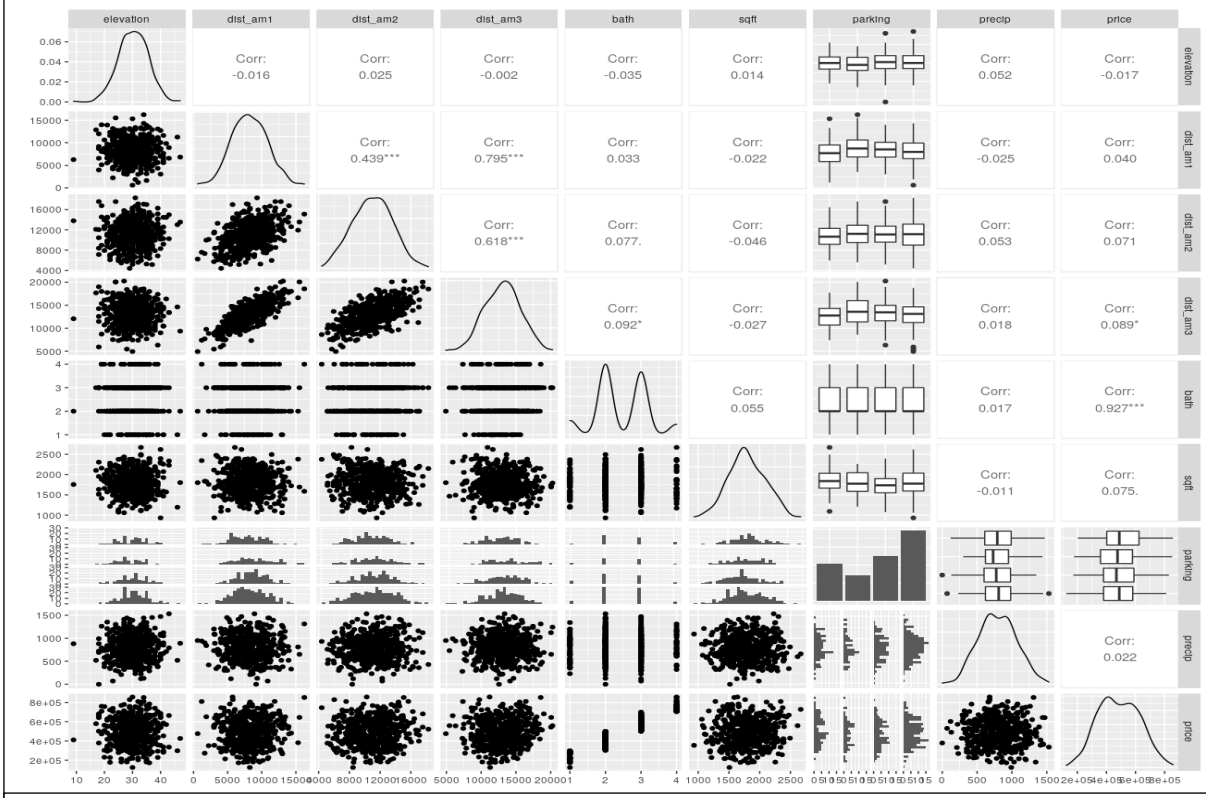
Were there any outliers present. If yes what methods did you use to find them.

Yes there are outliers. I used outlierTest to confirm what I was seeing in the graphs of ggpair and in the residual plots (autoplot- Cooks distance)

What methods did you use to choose the final model

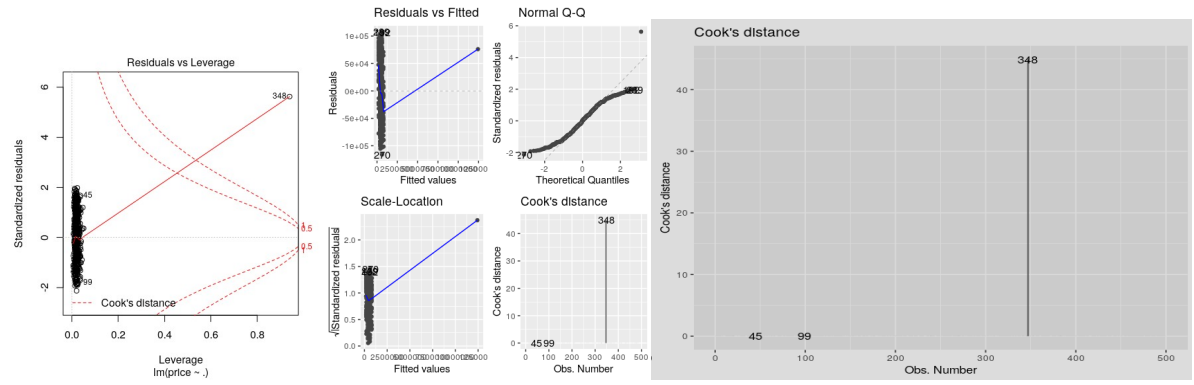
Since it has a large number of variables but not high enough to be unable to take all the combinations I took all the combinations

## Exploratory analysis:




I used ggpairs for exploratory analysis. It allows as to see all the scatterplots between the variables and barplots in categorical variables like parking. In the image above its the second ggpairs command I run after I removed an outlier which was obvious in the first ggpair plot.
The above graph with the outlier removed allows us to see the linear relationship between the dist_am1, dist_am2 and dist_am3 variables and give us an idea about the rest of the variables.
We see that parking is a categorical variable and bath is a discrete variable that could be considered categorical.

## Model Diagnostics and outlier detection




On model diagnostic I used summary to get an idea about the model. It had a really high R^2(adj). From the summary we get that the significant variables are bath and parking. Also from the residuals plots, scatterplots etc we can see that there is an outlier. I confirmed the outlier with outlierTest.

## Model selection

Since we have a big number of variables but not big enough not to be able to take all the combinations I choosed to take all possible combinations. (ols_step_all_possible()). Then I checked the best with each criteria in the variable model.selection and ended with the variables bath, sqft and parking. The differences within each critiria column are neglible.
Comparing the model (only bath) with the combination of bath with the others variables using anova doesnt seem to add usefull info to our model.
Looking the summary of the models we see that the model price~bath+parking parking breaks into dummy variable. One of them adds significant info to our model. Since the anova model doesnt show any real benefit from adding parking variable I decided to drop it.

The final model is
Price=58402 + 175635*bath

## Concluding remarks

From the above we see that our final model is price =58402+175635*bath which gives a high Adjusted R-squared 0.8586.
That means that we can predict the price of the house based on the number of baths. More particulary for every bath we add the price of the house increase by 175635