Text

Description automatically generatedInitial Data Summary

Data zip code: 98312

# of data points: 277

Average Price: $437,161

Average Bed: 3.253

Average Bath: 2.097

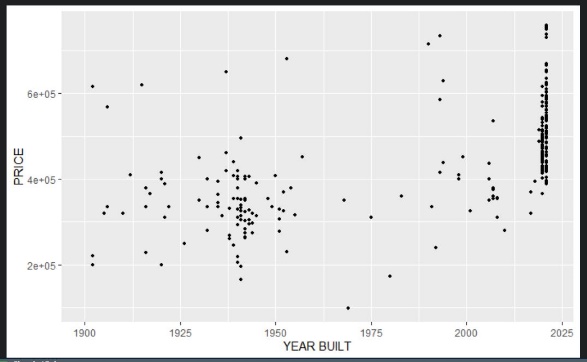
Average Square Feet: 1834

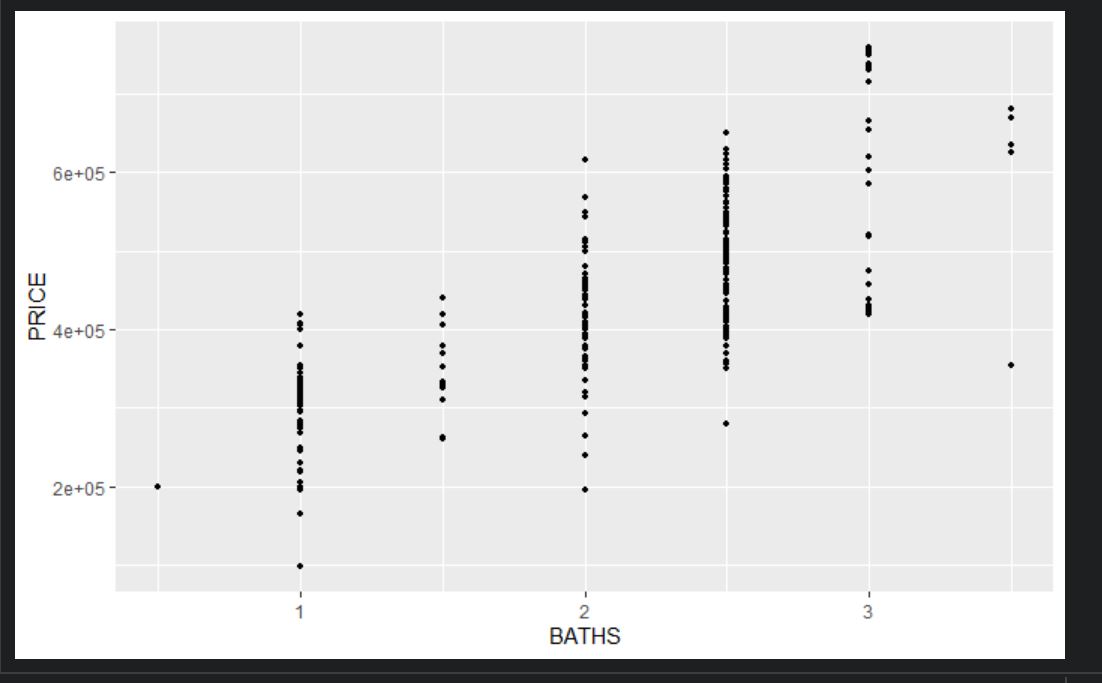
Average Lot Size: 6395sqft

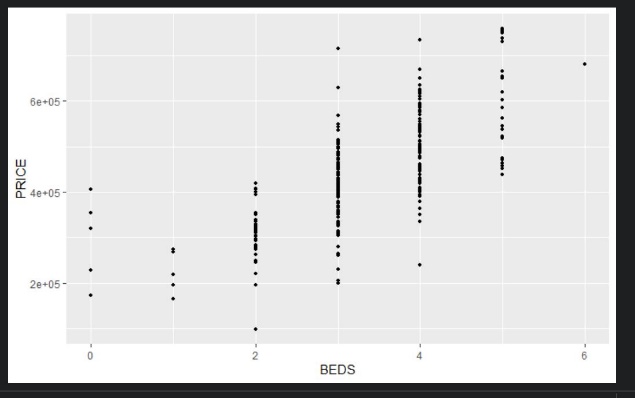
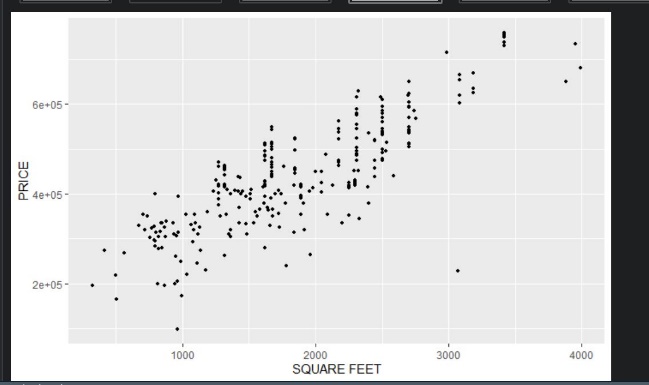
Average year built: 1988

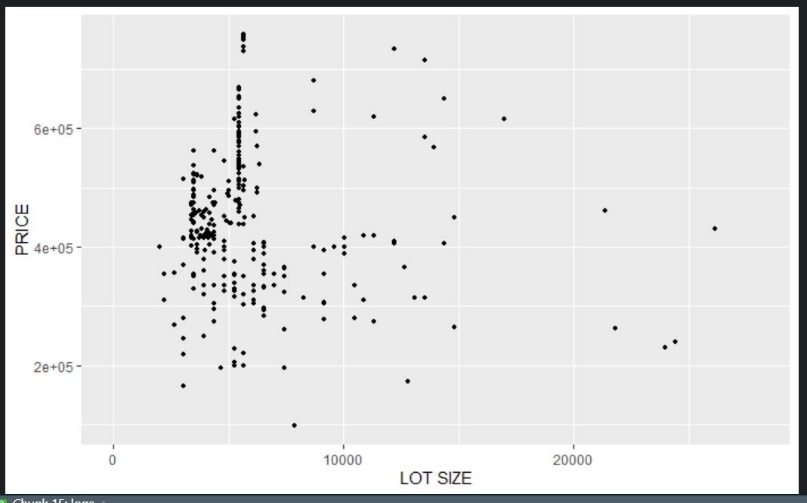
Average Days on Market: 158

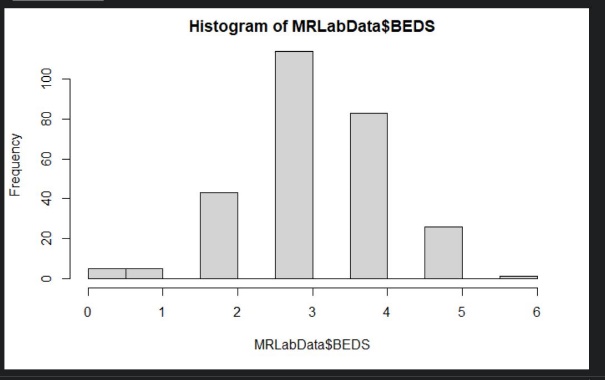
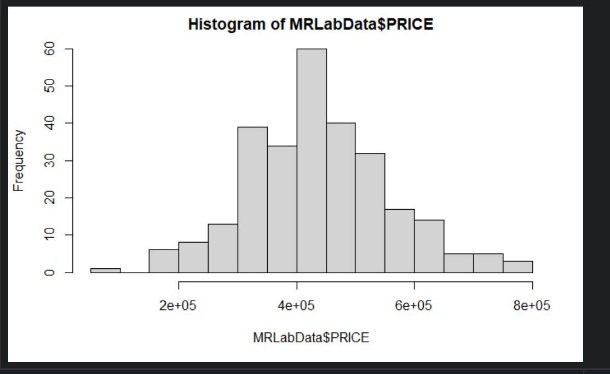
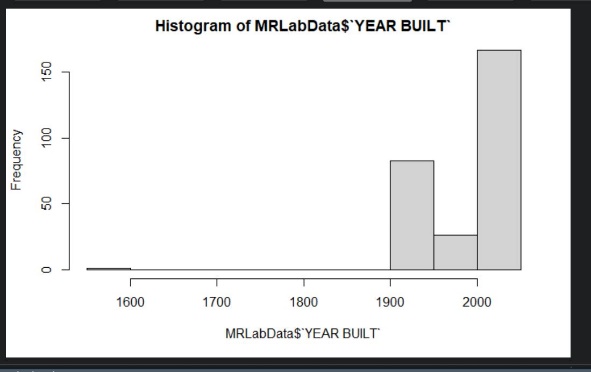
Average $/sqft: 258.6

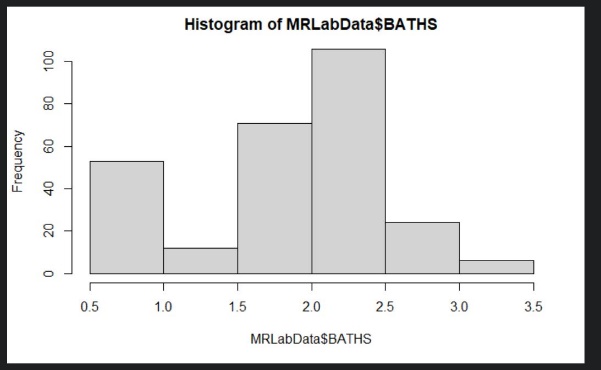
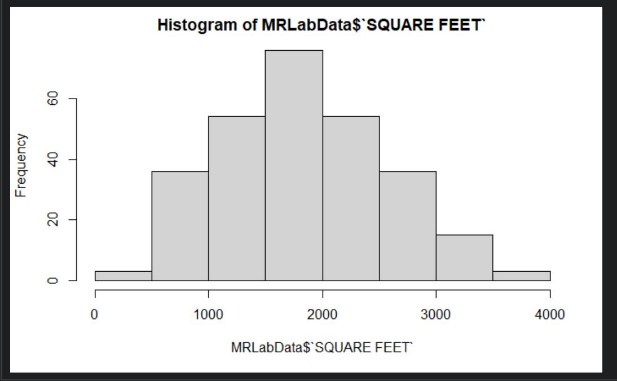
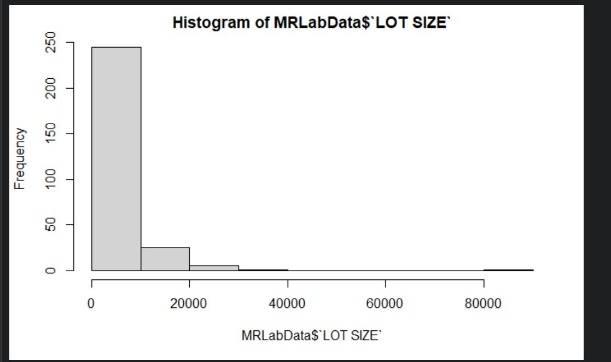


Ggplot scatterplots

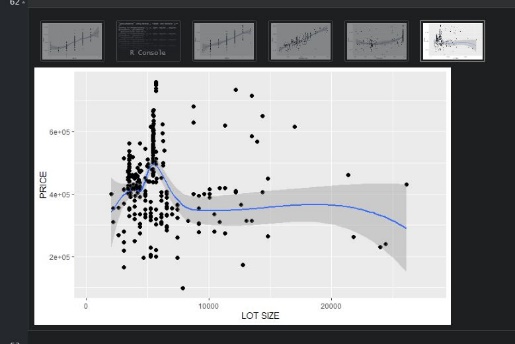


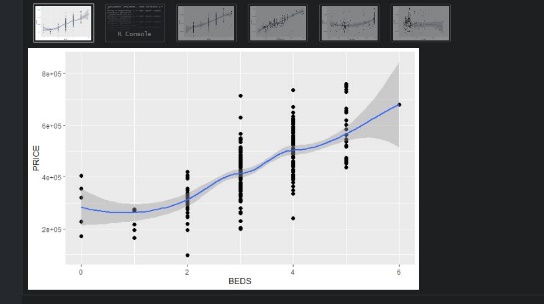


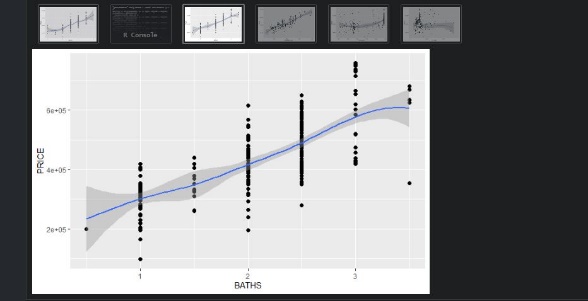
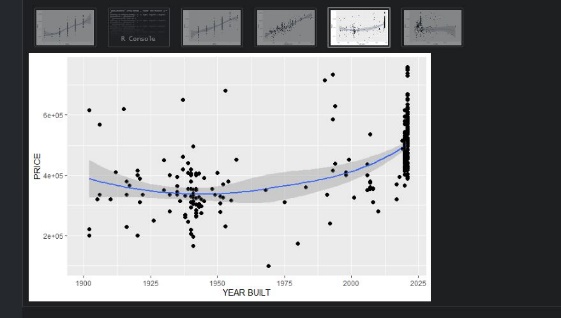
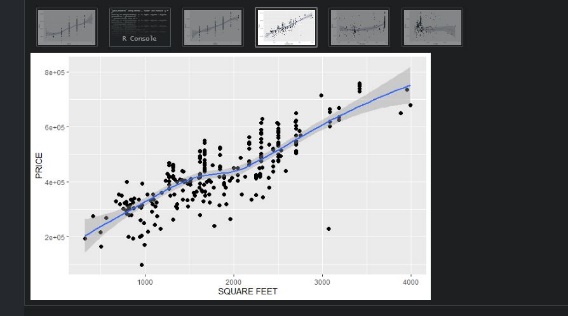
Histograms of initial data



From initial observation of the histograms and scatterplots it looks like square feet as the x variable (independent variable) and price as the y variable (dependent variable). best explain sales price. These graphs have a symmetric skew and form an even cluster around the linear relationship. The bath and bed variables would potentially not be the best dependent variables to utilize because their distribution is highly skewed.

The data can be utilized by data transformation and depict in our linear models would be square-feet, lot size, year built, beds, and price. OLS regression was performed below as seen.





When analyzing the graphs, each variable was plotted with the independent variable as price. With Lot Size you can see outliers and a skewed distribution. With beds, baths, years built, and square feet, all distributions seemed to relatively follow the linear coefficient and r^2 values, with minimal outliers shown.

Because regression 1 has the highest correlation between variables my analysis will be conducted on the image below. As shown below, when conducting a linear model and summary of the model, the variables with three stars show extremely high significance and support my hypothesis previously mentioned that square feet is the best conducive variable when determining price. Year built is also highly correlated to square feet showing great significance as well. The standard errors for each variable show minimal error with a t value that is significant. The remaining variables when all plotted together show minimal impact on the correlation coefficient (beds, baths, lotsize) but also have somewhat low p-values which suggest they are not significant but do play a part in shaping the correlation. To support this statement I ran another linear model with just square feet and year built to see if the fit would improve and it slightly lowered from .7522(regression1) to .7508(myownregression), showing these variables have minimum impact on the overall correlation.

Text

Description automatically generated

When the residuals graph is plotted and the quantile-quantile plot is shown for regression 1, you can see that the estimated responses along the y axis are and fitted values on the x axis show there is a positive relationship showing normality of the data. The residuals vs fitted plot shows that the variables are somewhat correlated and show little to no problems with the data. In other words, there is little heteroscedasticity, or a nonlinear relationship is not present between all the variables. The data shown below closely relates to the regression2 and regression3 models in the .rmd file provided.

Chart, scatter chart

Description automatically generatedChart, line chart

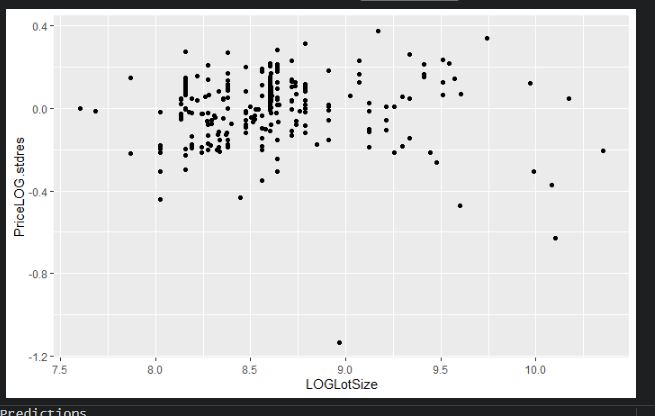
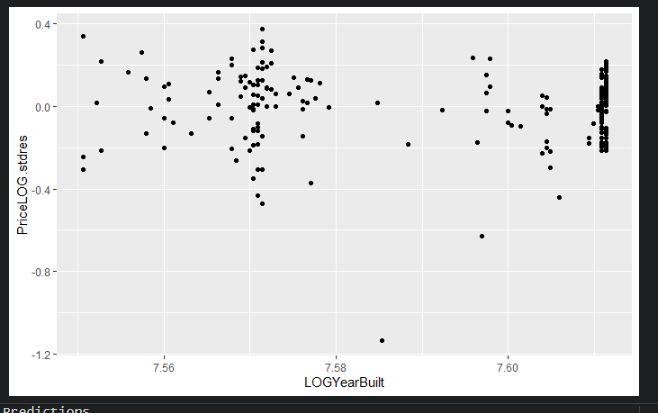
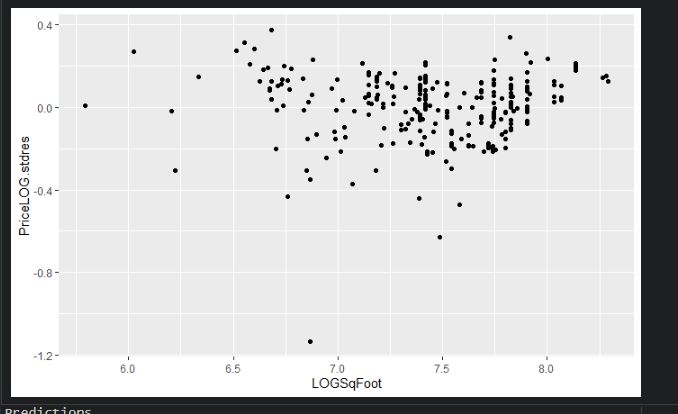
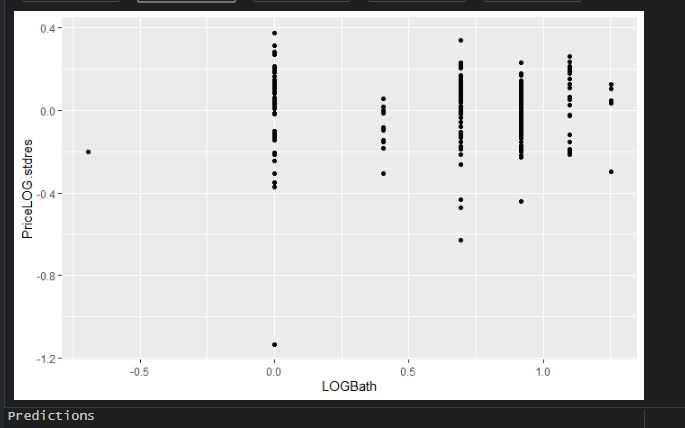
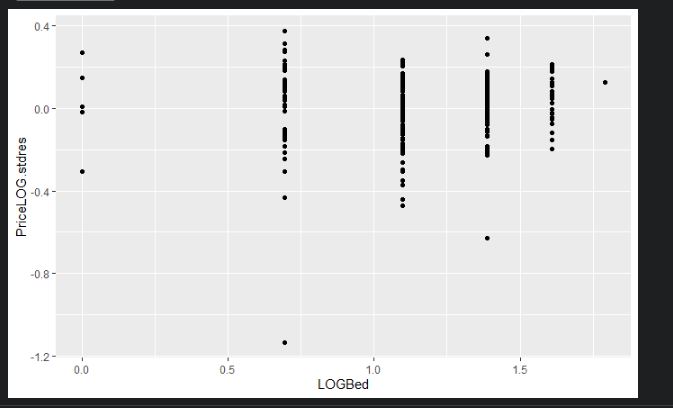
Description automatically generated

When analyzing the three different regression models, all three show very close correlation only having a .03 variance between their correlation coefficients. For the purposes of this assignment the regression model containing the highest correlation coefficient was taken which is regression1. To answer question #7, the answer is yes the model is significant.

Analyzing the f statistic of the logarithmic variables (the analysis of variance), we get 145 on 4, with 267 degrees of freedom. This statistic is utilized to determine the hypothesis for the relationship between the logarithmic variables and our independent variable price. Given a p-value of 2.2e-16, this evidence shows the regression model fits the data better than the model with no independent variables. In other words, the independent variable (price) improves the fit of the model and the overall correlation coefficient.

Text

Description automatically generated

Below are residual plots for each of the explanatory variables.

When understanding residuals in regression models, not all prediction models are accurate. An analysis of the residuals can offer a way to improve the regression model, and that’s what the plots above detail. Each of the variables above details using residual analysis with the independent variable as price and a brief analysis of each below:

Price -> Year Built & Square Feet & Lot Size – The residual model for years-built, square feet, and lot size shows lots of points on the right side with no u-shaped trend line or noticeable trends. The data seems to be un-clustered and shows minimal outliers. Overall, the fit of the residuals is not consistent with the highly correlated trend-line of year built.

Price -> Beds & Baths – The distribution of the residuals shows a consistency to the trend lines and correlation of the regression models. There are minimal outliers observed and the overall residual plots show the same linear traits of the correlation model.