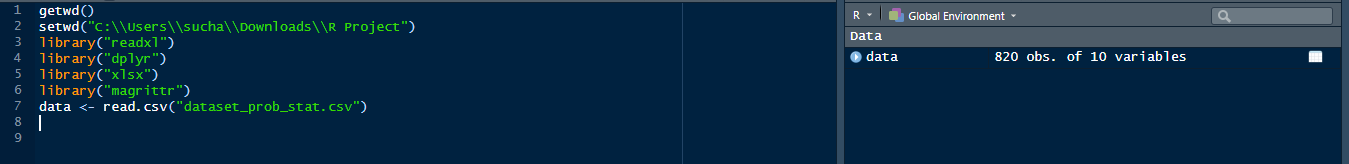
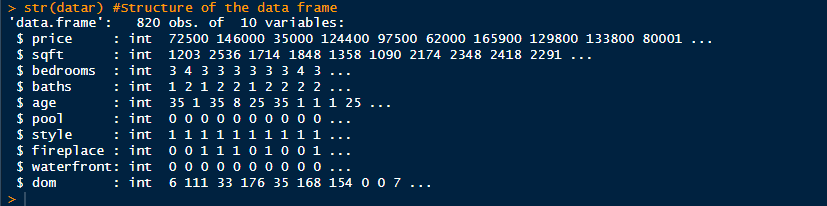
**House Prices in the USA**

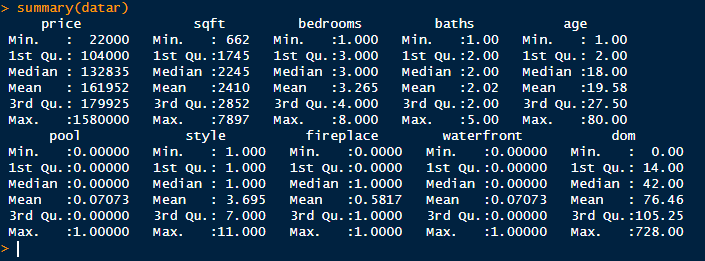
This data set contains house sale prices affecting house prices in the USA (Baton Rouge, Louisiana).

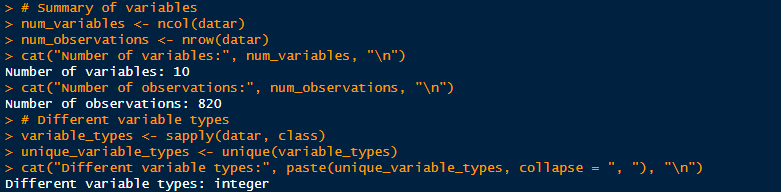
1. **Data exploration**

* *How many variables and observations are there in your data set?*



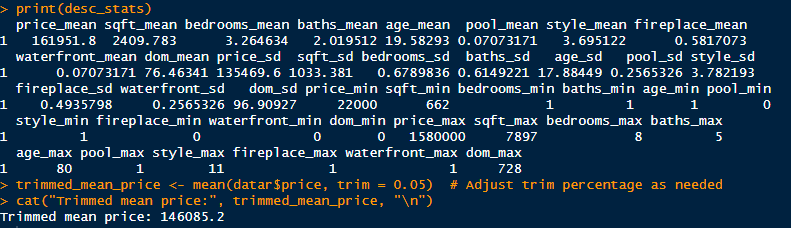






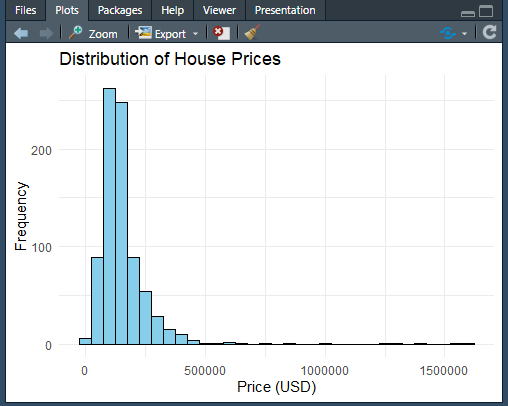
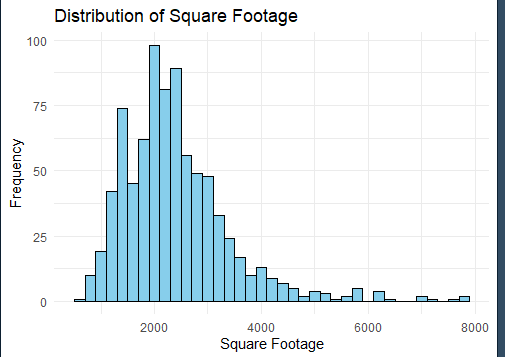
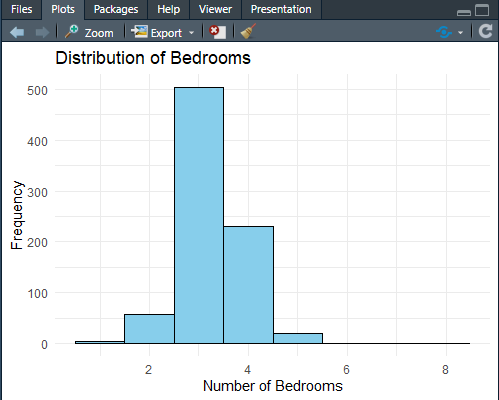
* *Descriptive statistics. Consider including the trimmed, if appropriate. Present your results clearly in a table and comment on your findings*.

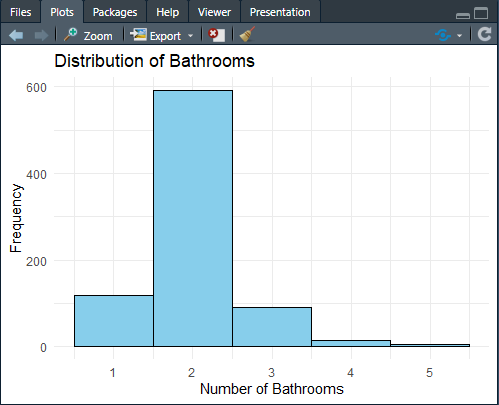
The trimmed mean of the price variable, calculated with a 5% trim, provides a robust measure of central tendency for house prices, accounting for potential outliers.



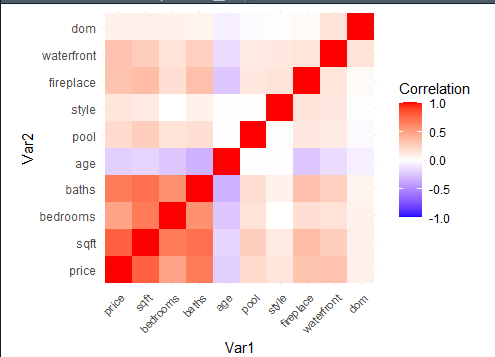
* + *Appropriate plots visualising the distributions of some of the variables. Comment on your findings.*

Create histograms for price, square footage, bedrooms, and baths

* + 1. 
    2. 
    3. 

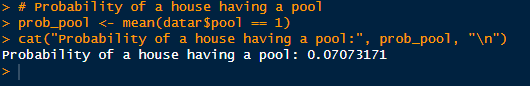


* + *Correlations between variables. Comment on your findings.*

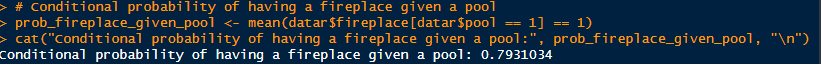


1. **Probability**
   * *A house is chosen at random from your data set. Compute:* ***–***

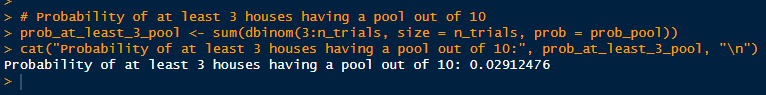
*the probability it has a pool.*



* + *the (conditional) probability it has a fireplace, given that it has a pool.*

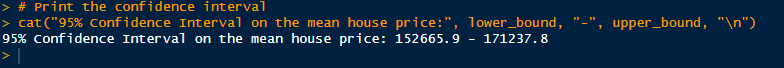


* + *Find the probability that, out of 10 houses chosen at random from your data set, at least 3 will have a pool. Explain your working.*

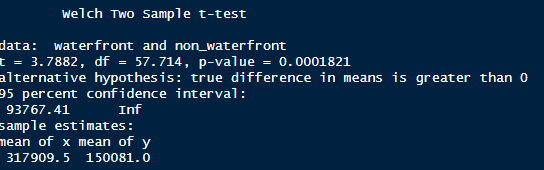
****

**Explanation:**

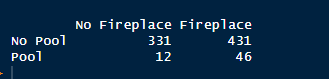
* + 1. We used the dbinom() function in R to calculate the probability of getting at least 3 houses with a pool out of 10, given the probability of a house having a pool (prob\_pool).
    2. The dbinom() function calculates the probability mass function of the binomial distribution, which gives the probability of obtaining a certain number of successes (houses with a pool) out of a fixed number of trials (houses chosen).
    3. We summed the probabilities for getting 3, 4, 5, ..., 10 houses with a pool to find the probability of at least 3 houses having a pool out of 10.
  + *Calculate a 95% confidence interval on the mean house price in the USA.*

****

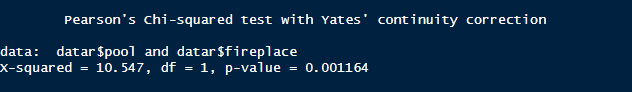
1. **Contingency tables and hypothesis tests**
   * *Test the hypothesis that the mean house price (over all house styles) is greater if a house is on the waterfront:*

**

* + *Create a contingency table showing relative frequencies for "Pool" and "No pool" according to whether a house has or hasn’t got a fireplace:*

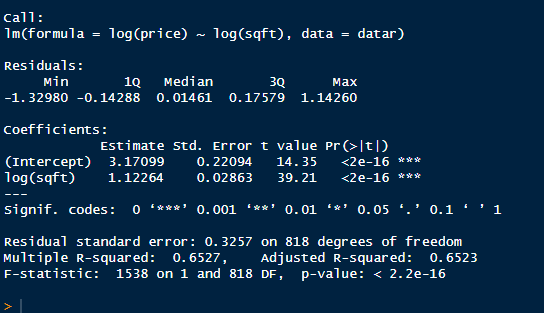
****

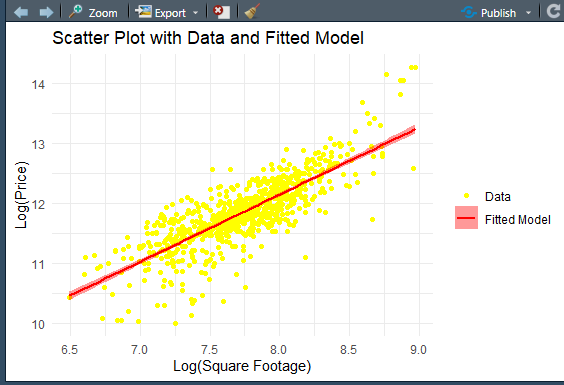
* + *Test whether a house having a fireplace is independent of whether it has a pool:*

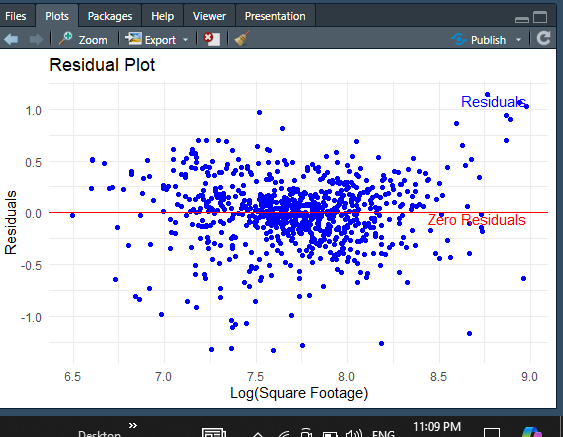
**

* + 1. Test used: Chi-squared test for independence.
    2. Null hypothesis (H0): Having a fireplace is independent of having a pool.
    3. Alternative hypothesis (H1): Having a fireplace is not independent of having a pool.

1. **Simple Linear Regression**

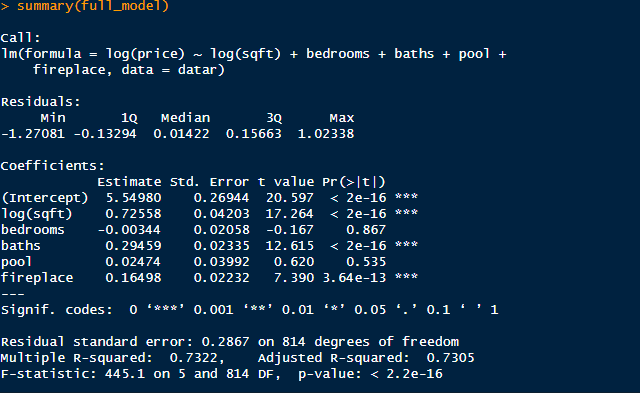
****

****

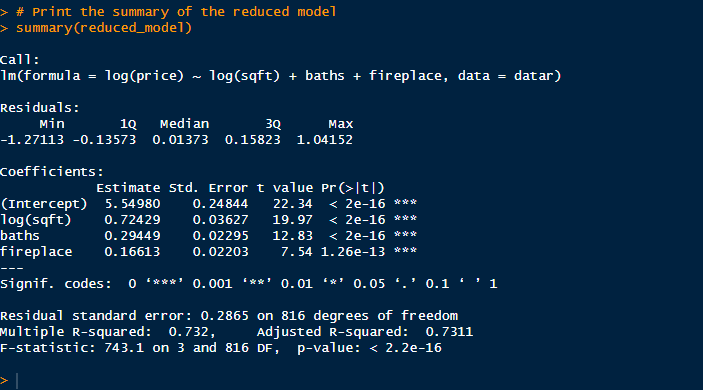
****

1. **Multiple Linear Regression**

* *Perform Multiple Linear Regression*

**

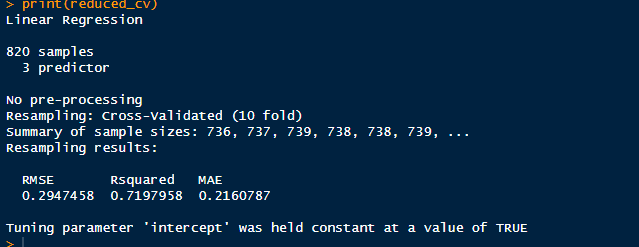
* *Feature Selection for Reduced Model*



**Justification:**

1. Backward elimination iteratively removes the least significant predictor variables from the full model until all remaining variables have significant p-values.
2. This method helps prevent overfitting by reducing the number of predictors and simplifying the model*.*

* *K-Fold Cross Validation to Compare Model Performances:*

****

**-0-**