Statistical analysis on Housing data

Elias Fedai

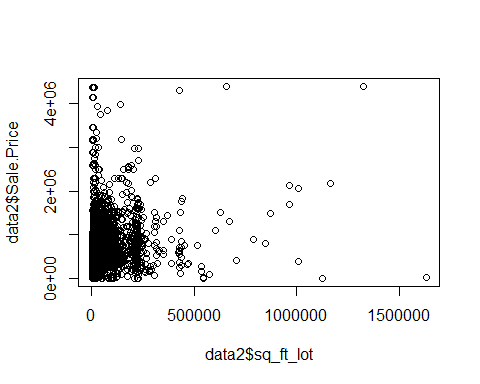
7/11/2019

## Task A

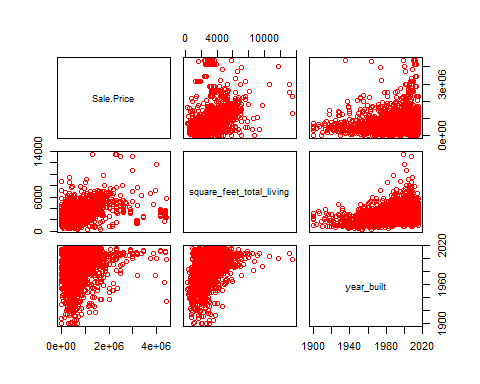
* The reason for removing data points are either there is a n/a in place of a value which can cause problems in R, or simple to remove unwanted data points that really serve no purpose in the problem you are looking to solve and can possible obscure or falsely change your results.

## Task B

## [1] 1



## [1] 2

 ## Task C

##   
## Call:  
## lm(formula = Sale.Price ~ sq\_ft\_lot, data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2016064 -194842 -63293 91565 3735109   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.418e+05 3.800e+03 168.90 <2e-16 \*\*\*  
## sq\_ft\_lot 8.510e-01 6.217e-02 13.69 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 401500 on 12863 degrees of freedom  
## Multiple R-squared: 0.01435, Adjusted R-squared: 0.01428   
## F-statistic: 187.3 on 1 and 12863 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = Sale.Price ~ square\_feet\_total\_living + year\_built,   
## data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1719467 -121308 -42621 44230 3916857   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.114e+06 3.808e+05 -13.43 <2e-16 \*\*\*  
## square\_feet\_total\_living 1.714e+02 3.346e+00 51.24 <2e-16 \*\*\*  
## year\_built 2.679e+03 1.923e+02 13.93 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 357500 on 12862 degrees of freedom  
## Multiple R-squared: 0.2184, Adjusted R-squared: 0.2183   
## F-statistic: 1797 on 2 and 12862 DF, p-value: < 2.2e-16

* The R2 and adjusted R2 on both models did not vary much, which would be indictive of a good model. The addition of additional predictors did slightly elevate the R2 bringing it a little closer to 1, but the residual error on both models is enormous and a little alarming. ## Task D
* The betas in the model i created are for the square\_feet\_total\_living the would be 171.4, and year\_built is 2679.0. These values represent the coefficient values.these values tell us about the relationship between sale price and each of the predictor variables.Also, the degree each predictor has on the affect of the outcome of the model.

## Task E

## 2.5 % 97.5 %  
## (Intercept) 6.343730e+05 6.492698e+05  
## sq\_ft\_lot 7.291208e-01 9.728641e-01

## 2.5 % 97.5 %  
## (Intercept) -5860366.2970 -4367673.7301  
## square\_feet\_total\_living 164.8777 177.9934  
## year\_built 2302.1248 3056.0179

* The quantiles or regression coefficient values are being shown at both a 2.5% and 97.5% confidence interval. On both occasions we never see a cross over 0, so these variables are significant.

## Task F

## Analysis of Variance Table  
##   
## Model 1: Sale.Price ~ sq\_ft\_lot  
## Model 2: Sale.Price ~ square\_feet\_total\_living + year\_built  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 12863 2.0734e+15   
## 2 12862 1.6441e+15 1 4.2931e+14 3358.7 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* based on the results on the analysis the increased value of F along with a significantly small value for Pr(>F) would indicate a signicant improved fit of the model.

## Task G

