HW3

Alan Jones

February 15, 2017

White\_wines<-read.csv("White\_wines.csv")

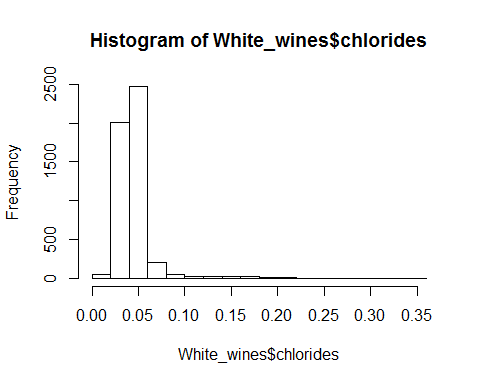
summary(White\_wines)

## fixed.acidity volatile.acidity citric.acid residual.sugar   
## Min. : 3.800 Min. :0.0800 Min. :0.0000 Min. : 0.600   
## 1st Qu.: 6.300 1st Qu.:0.2100 1st Qu.:0.2700 1st Qu.: 1.700   
## Median : 6.800 Median :0.2600 Median :0.3200 Median : 5.200   
## Mean : 6.855 Mean :0.2782 Mean :0.3342 Mean : 6.391   
## 3rd Qu.: 7.300 3rd Qu.:0.3200 3rd Qu.:0.3900 3rd Qu.: 9.900   
## Max. :14.200 Max. :1.1000 Max. :1.6600 Max. :65.800   
## chlorides free.sulfur.dioxide total.sulfur.dioxide  
## Min. :0.00900 Min. : 2.00 Min. : 9.0   
## 1st Qu.:0.03600 1st Qu.: 23.00 1st Qu.:108.0   
## Median :0.04300 Median : 34.00 Median :134.0   
## Mean :0.04577 Mean : 35.31 Mean :138.4   
## 3rd Qu.:0.05000 3rd Qu.: 46.00 3rd Qu.:167.0   
## Max. :0.34600 Max. :289.00 Max. :440.0   
## density pH sulphates alcohol   
## Min. :0.9871 Min. :2.720 Min. :0.2200 Min. : 8.00   
## 1st Qu.:0.9917 1st Qu.:3.090 1st Qu.:0.4100 1st Qu.: 9.50   
## Median :0.9937 Median :3.180 Median :0.4700 Median :10.40   
## Mean :0.9940 Mean :3.188 Mean :0.4898 Mean :10.51   
## 3rd Qu.:0.9961 3rd Qu.:3.280 3rd Qu.:0.5500 3rd Qu.:11.40   
## Max. :1.0390 Max. :3.820 Max. :1.0800 Max. :14.20   
## quality   
## Min. :3.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.878   
## 3rd Qu.:6.000   
## Max. :9.000

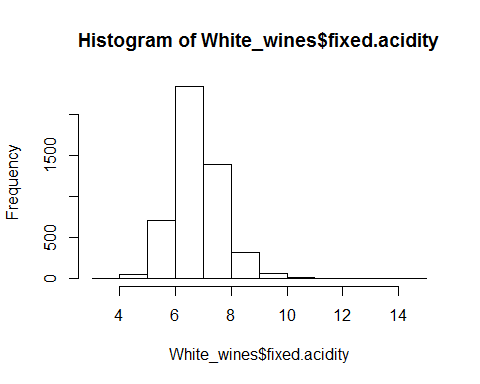
#examine each variable in a univariate graph  
variable.names(White\_wines)

## [1] "fixed.acidity" "volatile.acidity" "citric.acid"   
## [4] "residual.sugar" "chlorides" "free.sulfur.dioxide"   
## [7] "total.sulfur.dioxide" "density" "pH"   
## [10] "sulphates" "alcohol" "quality"

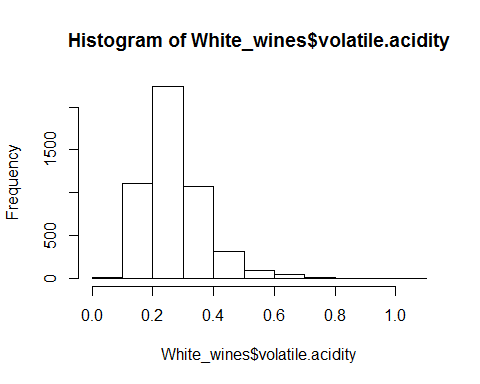
hist(White\_wines$chlorides)



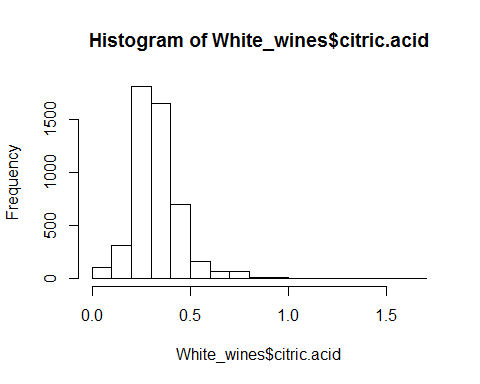
hist(White\_wines$fixed.acidity)



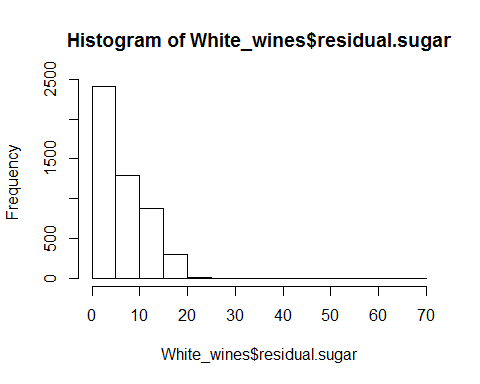
hist(White\_wines$volatile.acidity)



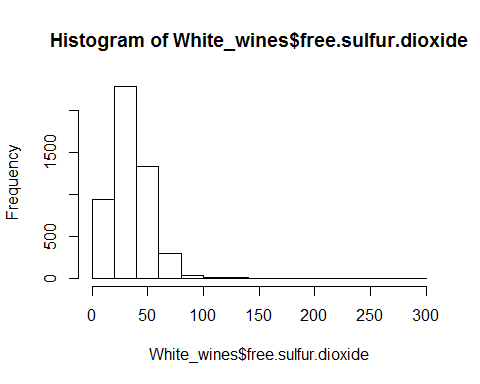
hist(White\_wines$citric.acid)



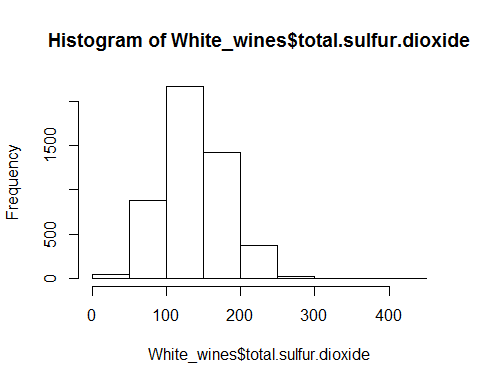
hist(White\_wines$residual.sugar)



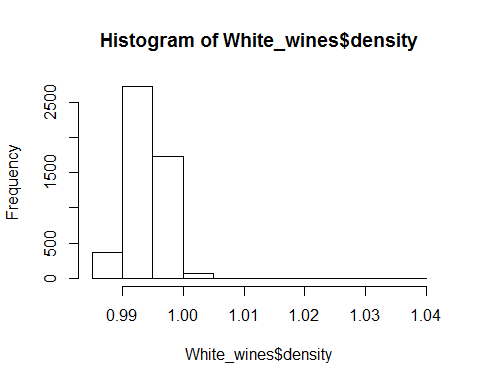
hist(White\_wines$free.sulfur.dioxide)



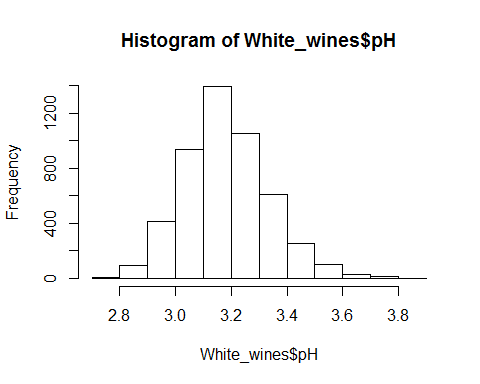
hist(White\_wines$total.sulfur.dioxide)



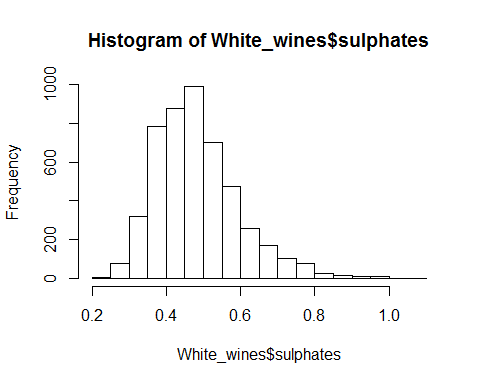
hist(White\_wines$density)



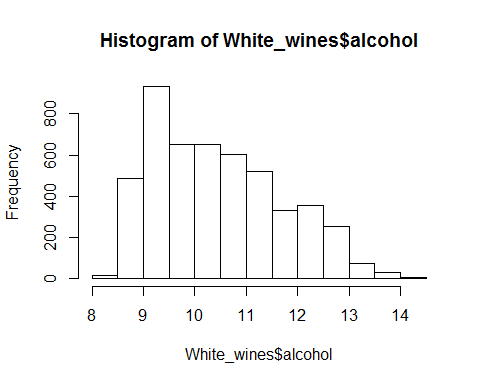
hist(White\_wines$pH)



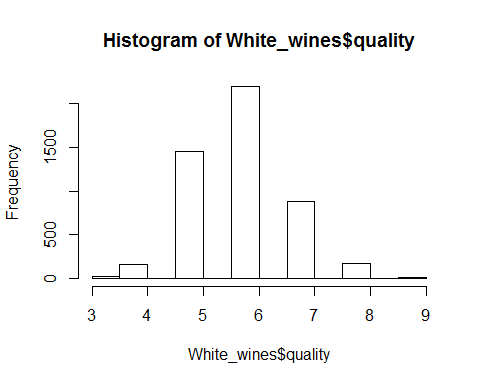
hist(White\_wines$sulphates)



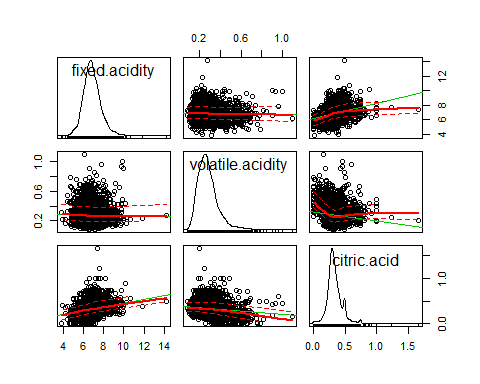
hist(White\_wines$alcohol)



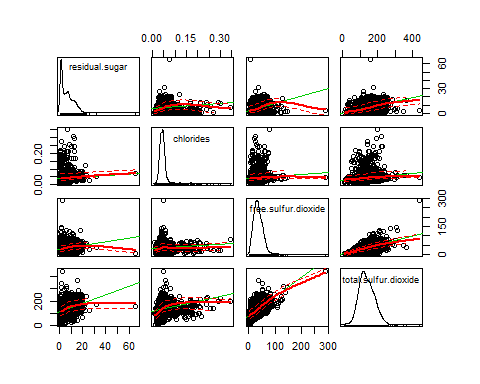
hist(White\_wines$quality)



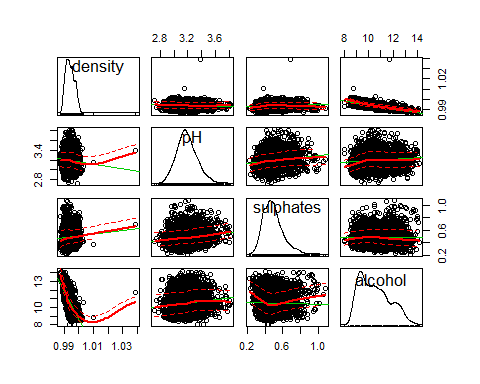
#To get multivariate graphs  
library(car)  
scatterplotMatrix(~fixed.acidity + volatile.acidity + citric.acid, data=White\_wines)



scatterplotMatrix(~residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide, data=White\_wines)



scatterplotMatrix(~density + pH + sulphates + alcohol, data=White\_wines)



#To examine regression  
RegModel.1 <-   
 lm(quality~alcohol+chlorides+citric.acid+density+fixed.acidity+free.sulfur.dioxide+pH+residual.sugar+sulphates+total.sulfur.dioxide+volatile.acidity,  
 data=White\_wines)  
summary(RegModel.1)

##   
## Call:  
## lm(formula = quality ~ alcohol + chlorides + citric.acid + density +   
## fixed.acidity + free.sulfur.dioxide + pH + residual.sugar +   
## sulphates + total.sulfur.dioxide + volatile.acidity, data = White\_wines)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8348 -0.4934 -0.0379 0.4637 3.1143   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.502e+02 1.880e+01 7.987 1.71e-15 \*\*\*  
## alcohol 1.935e-01 2.422e-02 7.988 1.70e-15 \*\*\*  
## chlorides -2.473e-01 5.465e-01 -0.452 0.65097   
## citric.acid 2.209e-02 9.577e-02 0.231 0.81759   
## density -1.503e+02 1.907e+01 -7.879 4.04e-15 \*\*\*  
## fixed.acidity 6.552e-02 2.087e-02 3.139 0.00171 \*\*   
## free.sulfur.dioxide 3.733e-03 8.441e-04 4.422 9.99e-06 \*\*\*  
## pH 6.863e-01 1.054e-01 6.513 8.10e-11 \*\*\*  
## residual.sugar 8.148e-02 7.527e-03 10.825 < 2e-16 \*\*\*  
## sulphates 6.315e-01 1.004e-01 6.291 3.44e-10 \*\*\*  
## total.sulfur.dioxide -2.857e-04 3.781e-04 -0.756 0.44979   
## volatile.acidity -1.863e+00 1.138e-01 -16.373 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7514 on 4886 degrees of freedom  
## Multiple R-squared: 0.2819, Adjusted R-squared: 0.2803   
## F-statistic: 174.3 on 11 and 4886 DF, p-value: < 2.2e-16

#To further examine regression  
RegModel.1 <-   
 lm(quality~alcohol+density+fixed.acidity+free.sulfur.dioxide+pH+residual.sugar+sulphates+volatile.acidity,  
 data=White\_wines)  
summary(RegModel.1)

##   
## Call:  
## lm(formula = quality ~ alcohol + density + fixed.acidity + free.sulfur.dioxide +   
## pH + residual.sugar + sulphates + volatile.acidity, data = White\_wines)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8246 -0.4938 -0.0396 0.4660 3.1208   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.541e+02 1.810e+01 8.514 < 2e-16 \*\*\*  
## alcohol 1.932e-01 2.408e-02 8.021 1.31e-15 \*\*\*  
## density -1.543e+02 1.834e+01 -8.411 < 2e-16 \*\*\*  
## fixed.acidity 6.810e-02 2.043e-02 3.333 0.000864 \*\*\*  
## free.sulfur.dioxide 3.349e-03 6.766e-04 4.950 7.67e-07 \*\*\*  
## pH 6.942e-01 1.034e-01 6.717 2.07e-11 \*\*\*  
## residual.sugar 8.285e-02 7.287e-03 11.370 < 2e-16 \*\*\*  
## sulphates 6.285e-01 9.997e-02 6.287 3.52e-10 \*\*\*  
## volatile.acidity -1.888e+00 1.095e-01 -17.242 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7512 on 4889 degrees of freedom  
## Multiple R-squared: 0.2818, Adjusted R-squared: 0.2806   
## F-statistic: 239.7 on 8 and 4889 DF, p-value: < 2.2e-16

You are now the data slave to the principal investigator Dr. Vinca Monster. Dr. M is in the Grape Program at State U, and you are just a poor graduate student trying to get your degree. Dr. M is interested in wine preferences and the influences of physico-chemical properties on preferences. Her laboratory has gathered an extensive dataset on Portugese white varietals.

You will find the white\_wines.csv file and its description in my github repo (<https://github.com/vhertzb/Regression-1>)[<https://github.com/vhertzb/Regression-1>].

Please use the techniques you have learned in the last two classes, specifically exploratory data analysis and linear regression, to determine association of the wine properties on preference.

Prepare a report for presentation at the next Monster lab meeting about this dataset.

Rubric:

Exploration (summary statistics (the m's), univariate graphs, multivariate graphs) Regression (Models explored, diagnostics completed, final model choice, justification)

Please include a concluding paragraph (or two) about the implications of your findings.