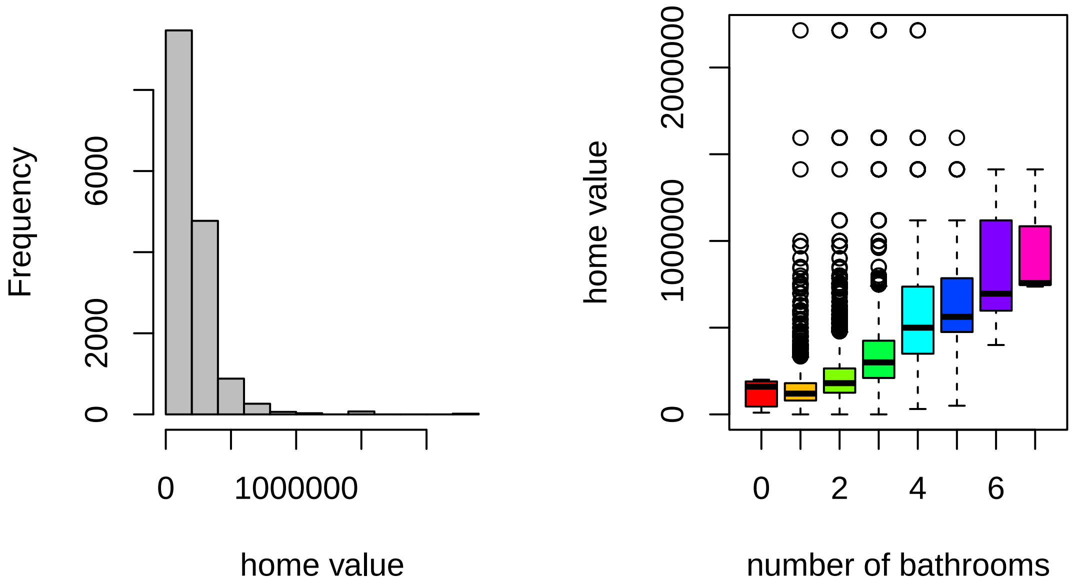
Homework 2 Assignment

This assignment contains TWO data analyses. Each question is 10 points.

Predicting House Prices

```
## Read in the data

homes <- read.csv("homes2004.csv")

# conditional vs marginal value


par(mfrow=c(1,2)) # 1 row, 2 columns of plots

hist(homes$VALUE, col="grey", xlab="home value", main="")

plot(VALUE ~ factor(BATHS),
    col=rainbow(8), data=homes[homes$BATHS<8,],
    xlab="number of bathrooms", ylab="home value")
```



```
# You can try some quick plots.  Do more to build your intuition!

#par(mfrow=c(1,2))
#plot(VALUE ~ STATE, data=homes,
#   col=rainbow(nlevels(homes$STATE)),
#   ylim=c(0,10^6), cex.axis=.65)
#plot(gt20dwn ~ FRSTHO, data=homes,
#   col=c(1,3), xlab="Buyer's First Home?",
#   ylab="Greater than 20% down")
```

Question 1

Regress log price onto all variables but mortgage. What is the R2? How many coefficients are used in this model and how many are significant at 10% FDR? Re-run regression with only the significant covariates, and compare R2 to the full model.

```
library(knitr) # library for nice R markdown output
# regress log(PRICE) on everything except AMMORT

pricey <- glm(log(LPRICE) ~ .-AMMORT, data=homes)

# extract pvalues

pvals <- summary(pricey)$coef[-1,4]

# example: those variable insignificant at alpha=0.05

names(pvals)[pvals>.05]

# you'll want to replace .05 with your FDR cutoff
# you can use the `-AMMORT' type syntax to drop variables
```

Question 2

Fit a regression for whether the buyer had more than 20 percent down (onto everything but AMMORT and LPRICE). Interpret effects for Pennsylvania state, 1st home buyers and the number of bathrooms. Add and describe an interaction between 1st home-buyers and the number of baths.

```
# create a var for downpayment being greater than 20%
homes$gt20dwn <-
  factor(0.2<(homes$LPRICE-homes$AMMORT)/homes$LPRICE)
```

Question 3

Focus only on a subset of homes worth > 100k. Train the full model from Question 1 on this subset. Predict the left-out homes using this model. What is the out-of-sample fit (i.e. R^2)? Explain why you get this value.

```
subset <- which(homes$VALUE>100000)

# Use the code ``deviance.R" to compute OOS deviance

source("deviance.R")

# Null model has just one mean parameter

ybar <- mean(log(homes$LPRICE[-subset]))

D0 <- deviance(y=log(homes$LPRICE[-subset]), pred=ybar)

# - don't forget family="binomial"!
# - use +A*B in forumula to add A interacting with B
```

Amazon Reviews

We will use the same datasets (review_subset.csv, word_freq.csv and words.csv) as in Assignment 1.

```r
data<-read.table("Review_subset.csv",header=TRUE)

words<-read.table("words.csv")
words<-words[,1]

doc_word<-read.table("word_freq.csv")
names(doc_word)<-c("Review ID","Word ID","Times Word" )
```

Question 4

We want to build a predictor of customer ratings from product reviews and product attributes. For these questions, you will fit a LASSO path of logistic regression using a binary outcome:

$$Y = 1 \quad \text{for 5 stars} \tag{1}$$
$$Y = 0 \quad \text{for less than 5 stars.} \tag{2}$$

Fit a LASSO model with only product categories. The start code prepares a sparse design matrix of 142 product categories. What is the in-sample R2 for the AICc slice of the LASSO path? Why did we use standardize FALSE?

```r
# Let's define the binary outcome

# Y=1 if the rating was 5 stars

# Y=0 otherwise

Y<-as.numeric(data$Score==5)

# (a) Use only product category as a predictor

library(gamlr)
```

```
## Loading required package: Matrix
```

```r
source("naref.R")

# Cast the product category as a factor
data$Prod_Category<-as.factor(data$Prod_Category)

class(data$Prod_Category)
```

[1] “factor”

```r
# Since product category is a factor, we want to relevel it for the LASSO.
# We want each coefficient to be an intercept for each factor level rather than a contrast.
# Check the extra slides at the end of the lecture.
# look inside naref.R. This function relevels the factors for us.

data$Prod_Category<-naref(data$Prod_Category)

# Create a design matrix using only products
```

```r
products<-data.frame(data$Prod_Category)

x_cat<-sparse.model.matrix(~., data=products)[,-1]

# Sparse matrix, storing 0's as .'s
# Remember that we removed intercept so that each category
# is standalone, not a contrast relative to the baseline category

colnames(x_cat)<-levels(data$Prod_Category)[-1]

# let's call the columns of the sparse design matrix as the product categories

# Let's fit the LASSO with just the product categories

lasso1<- gamlr(x_cat,   y=Y, standardize=FALSE,family="binomial",
lambda.min.ratio=1e-3)
```

Question 5

Fit a LASSO model with both product categories and the review content (i.e. the frequency of occurrence of words). Use AICc to select lambda. How many words were selected as predictive of a 5 star review? Which 10 words have the most positive effect on odds of a 5 star review? What is the interpretation of the coefficient for the word 'discount'?

```r
# Fit a LASSO with all 142 product categories and 1125 words

spm<-sparseMatrix(i=doc_word[,1],
                  j=doc_word[,2],
                  x=doc_word[,3],
                  dimnames=list(id=1:nrow(data),
                  words=words))

dim(spm) # 13319 reviews using 1125 words
```

```
[1] 13319 1125
```

```r
x_cat2<-cbind(x_cat,spm)

lasso2 <- gamlr(x_cat2, y=Y,lambda.min.ratio=1e-3,family="binomial")
```

Question 6

Continue with the model from Question 5. Run cross-validation to obtain the best lambda value that minimizes OOS deviance. How many coefficients are nonzero then? How many are nonzero under the 1se rule?

```r
cv.fit <- cv.gamlr(x_cat2,
                   y=Y,
                   lambda.min.ratio=1e-3,
                   family="binomial",
                   verb=TRUE)
```

```
fold 1,2,3,4,5,done.
```