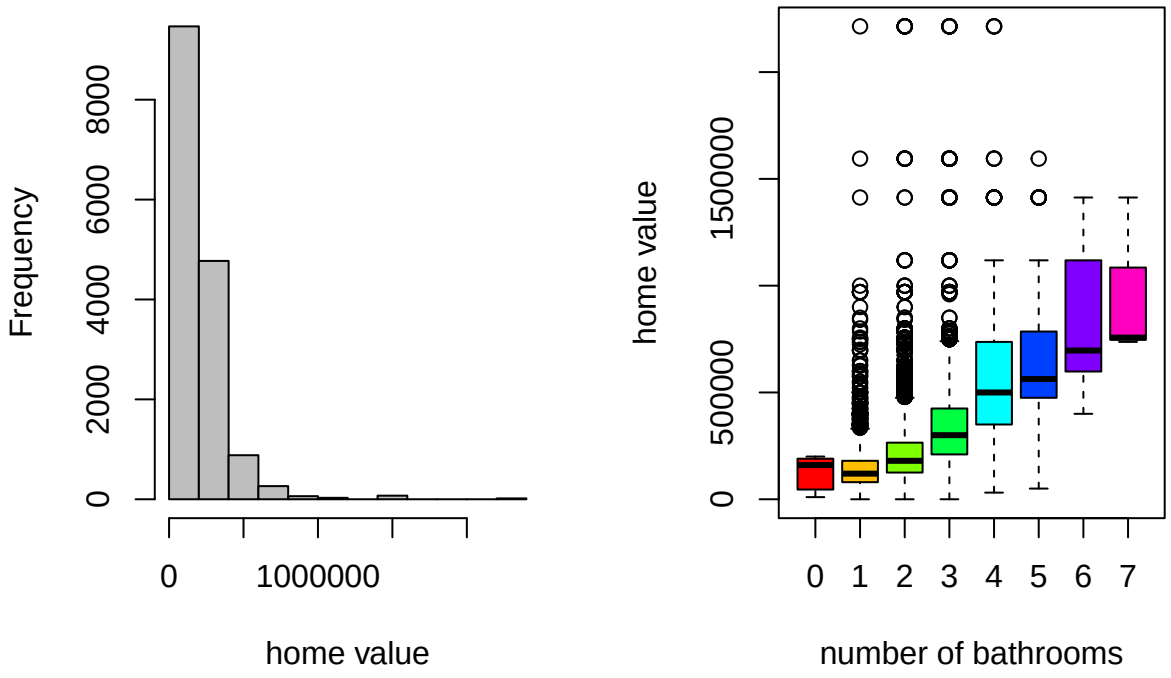
BUS 41201 Homework 2 Assignment

Shihan Ban, Yi Cao, Shri Lekkala, Ningxin Zhang

2 April 2024

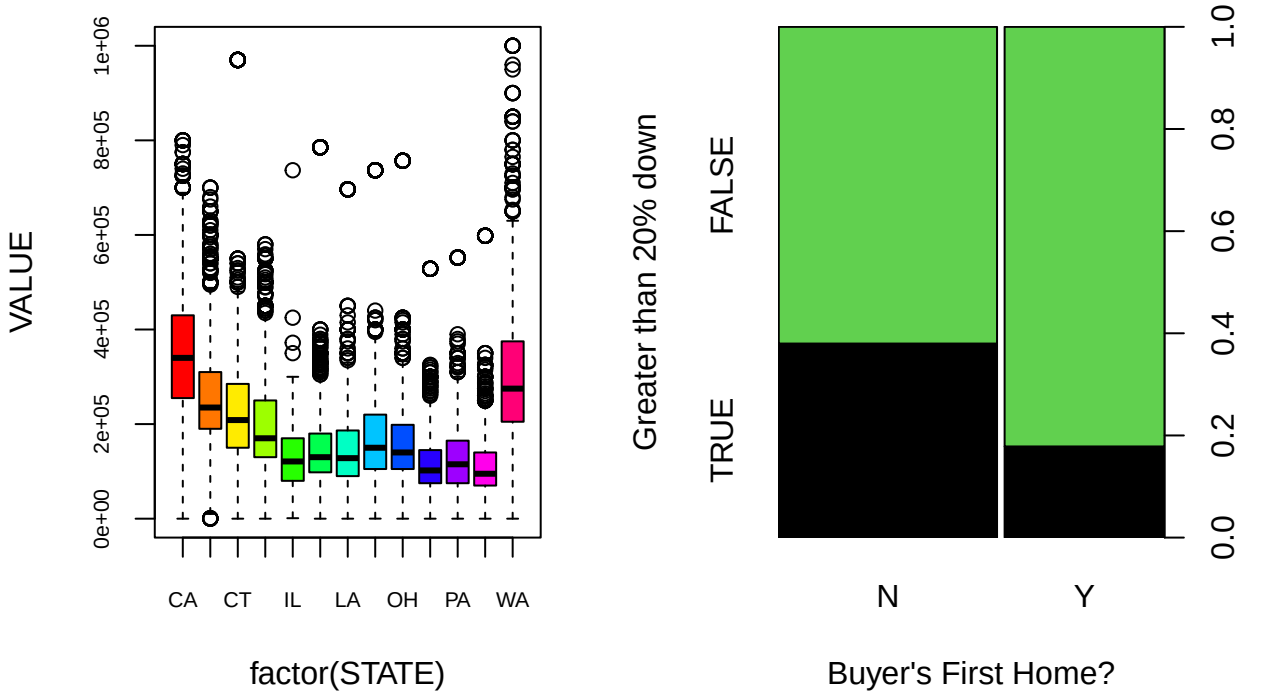
Setup

```
library(knitr) # library for markdown output

##### ******** Mortgage and Home Sales Data ******** #####

## 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")
```




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

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

par(mfrow=c(1,2))
plot(VALUE ~ factor(STATE), data=homes,
     col=rainbow(nlevels(factor(homes$STATE))),
     ylim=c(0,10^6), cex.axis=.65)
plot(gt20dwn ~ factor(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. (2 points)

```
# 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]
```

```
## [1] "ETRANSY" "NUNITS"  "STATECO" "STATECT" "BEDRMS"
```

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

Question 2

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

So there are 461 tests significant at the alpha level 0.05 and 348 tests significant at the alpha level 0.01.

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. R2)?
Explain why you get this value. (1 point)

```
# this is your training sample
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)
```

So the p-value cutoff for 1% FDR is: 0.002413249