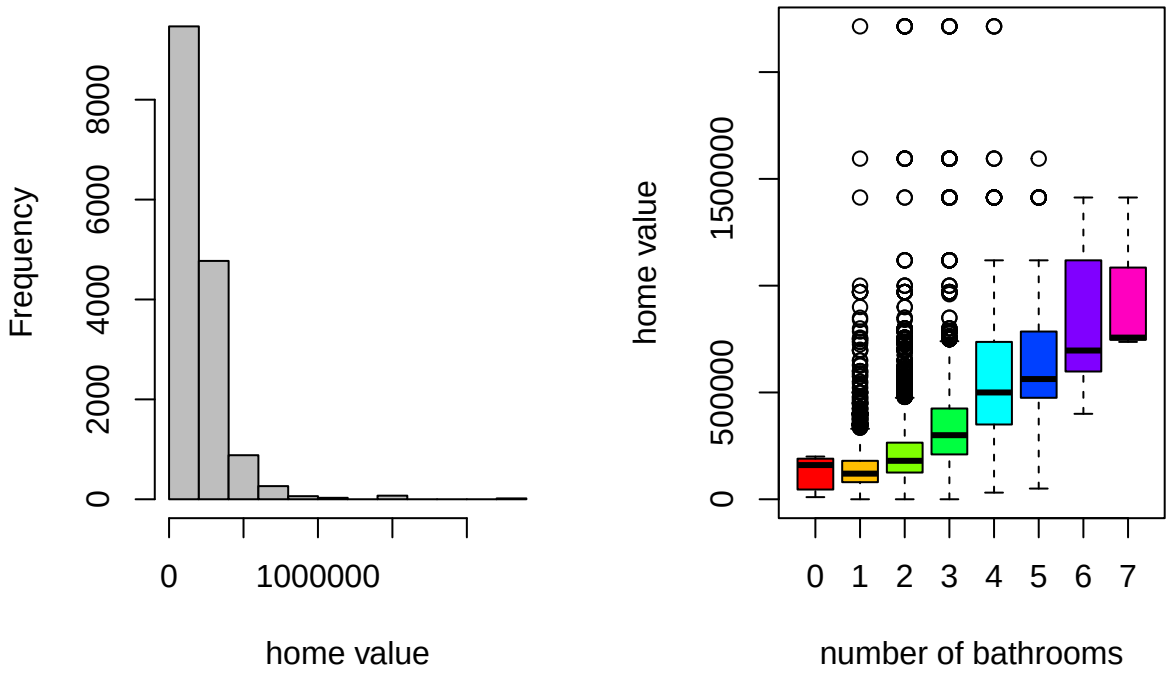
BUS 41201 Homework 2 Assignment

Shihan Ban, Yi Cao, Shri Lekkala, Ningxin Zhang

2 April 2024

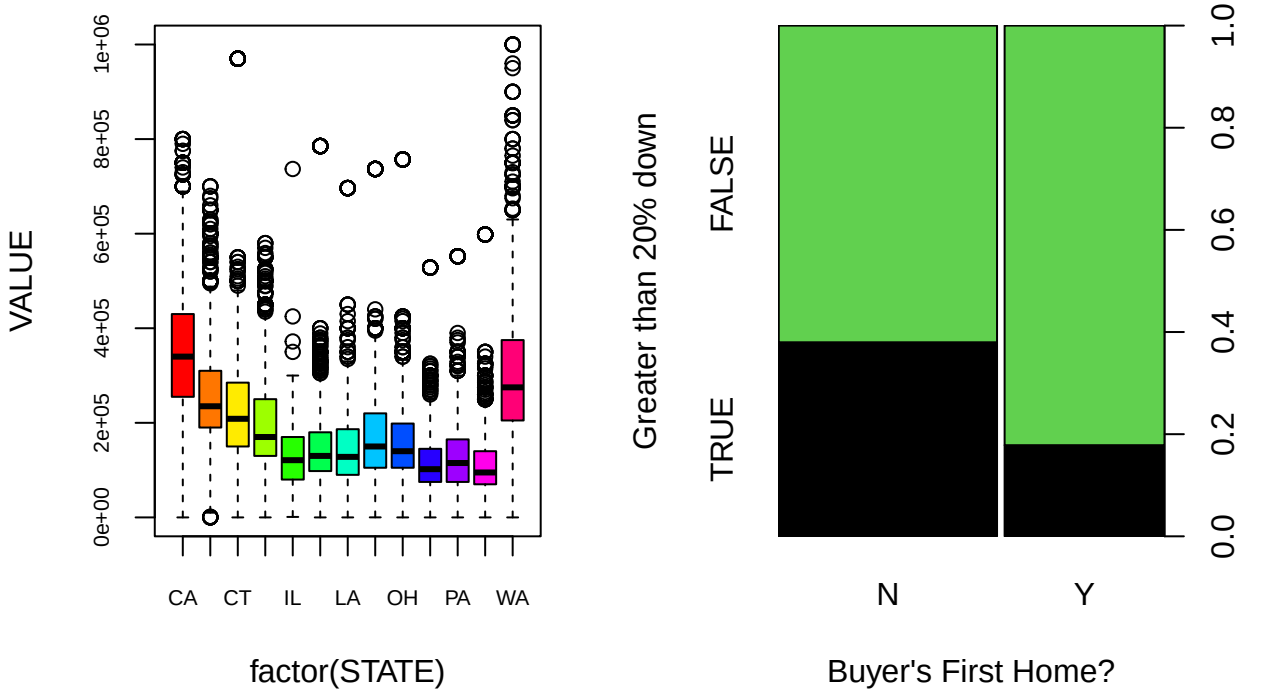
Setup

```
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 R-squared value the summary
summary_pricey <- summary(pricey)
R2 = 1 - summary_pricey$deviance / summary_pricey$null.deviance
R2
```

```
## [1] 0.4565419
```

So the R2 score is 0.4565419.

```r
# extract pvalues
pvals <- summary(pricey)$coef[-1,4]
length(pvals)
```

```
## [1] 42
```

So there are 42 coefficients in this model.

```r
# Find the p-value cutoff at the 10% FDR level

# To find the p-value cut off we first order the p values
pvals_ordered <- pvals[order(pvals, decreasing=F)]

# Next we use the function fdr_cut function defined in class class to find the cutoff at level 0.1
fdr_cut <- function(pvals, q){
  pvals <- pvals[!is.na(pvals)]
  N <- length(pvals)
  k <- rank(pvals, ties.method="min")
  alpha <- max(pvals[ pvals<= (q*k/N) ])
  return(alpha)
}

p_cutoff = fdr_cut(pvals_ordered, q=0.1)
p_cutoff
```

```
## [1] 0.03792594
```

```r
# Find the number of significant coefficients at this level
sum(pvals < p_cutoff)
```

```
## [1] 36
```

So out of the 42 coefficients, 36 are significant at the 10% FDR level.

```r
# Extract significant coefficients
significant_vars = names(pvals)[pvals < p_cutoff]

# Manually construct the formula string
response_var = "log(LPRICE)"
covariates = paste(significant_vars, collapse = " + ")
reduced_formula_str = paste(response_var, "~", covariates)

# # Rerun the regression with the significant covariates
# reduced_model = glm(reduced_formula_str, data=homes)
#
# # Extract R-squared value the summary
# summary_reduced_model = summary(reduced_model)
# R2_reduced = 1 - summary_reduced_model$deviance / summary_reduced_model$null.deviance
# R2_reduced
```

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. (2 points)

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

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