GR5206 Homework 2

Hanao Li

9/21/2018

Part 1: Loading and Cleaning the Data in R

i.

```
# Load data into housing dataframe
setwd('~/Desktop/R')
housing <- read.csv('NYChousing.csv', header = TRUE)
```

ii.

```
# Check rows and columns of the dataframe
dim(housing)
```

```
## [1] 2506   22
```

```
n <- nrow(housing)  # store rows to solve question no.5
```

There are 2506 rows and 22 columns

iii.

```
# Check the number of missing rows for each column variable
apply(is.na(housing), 2, sum)
```

```
##                          UID                 PropertyName
##                            0                            0
##                          Lon                          Lat
##                           15                           15
##                     AgencyID                         Name
##                            0                            0
##                        Value                      Address
##                           52                            0
##               Violations2010                   REACNumber
##                            0                         1873
##                     Borough                           CD
##                            0                            0
##          CityCouncilDistrict                  CensusTract
##                           10                            0
##                BuildingCount                    UnitCount
##                            0                            0
##                    YearBuilt                        Owner
##                            0                            0
##                  Rental.Coop            OwnerProfitStatus
##                            0                            0
##     AffordabilityRestrictions StartAffordabilityRestrictions
##                            0                            5
```

This command check the total missing rows of each column variable, for example there are 52 missing data for ‘Value’ variable

iv.

```
# Remove the missing rows for Value
housing <- housing[!is.na(housing$Value),]
```

v.

```
# Check how many rows were removed
n - nrow(housing)
```

```
## [1] 52
```

From the result, we could see that there are 52 rows were removed and this agress with the result we get from (iii)

vi.

```
# Create new variable called logValue
housing['logValue'] <- log(housing$Value)
summary(housing$logValue)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.41   12.49   13.75   13.68   14.80   20.47
```

The minimum is 8.41, the median is 13.75, the mean is 13.68, the maximum is 20.47

vii.

```
# Create new variable called logUnits
housing['logUnits'] <- log(housing$UnitCount)
summary(housing$logUnits)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.773   3.892   3.775   4.691   9.640
```

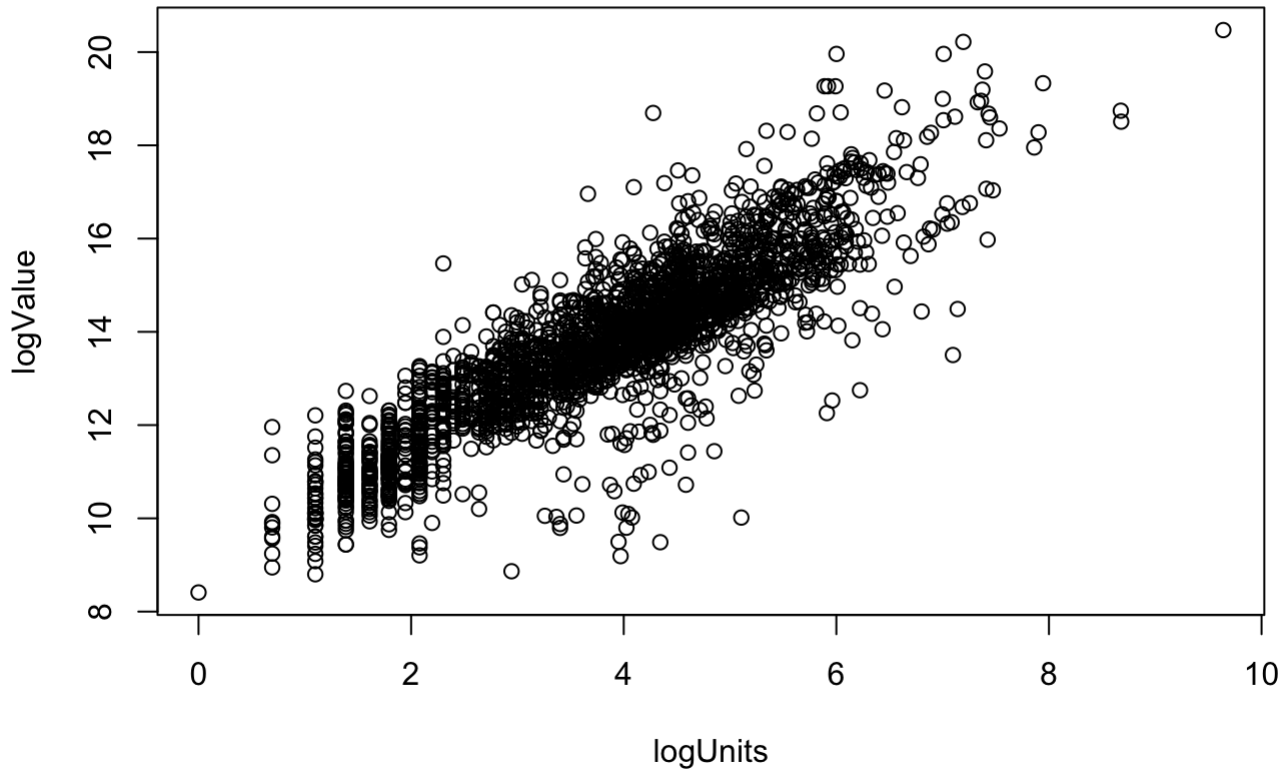
viii.

```
# Create new variable called after1950
housing['after1950'] <- ifelse(housing$YearBuilt >= 1950, TRUE, FALSE)
```

Part 2: EDA

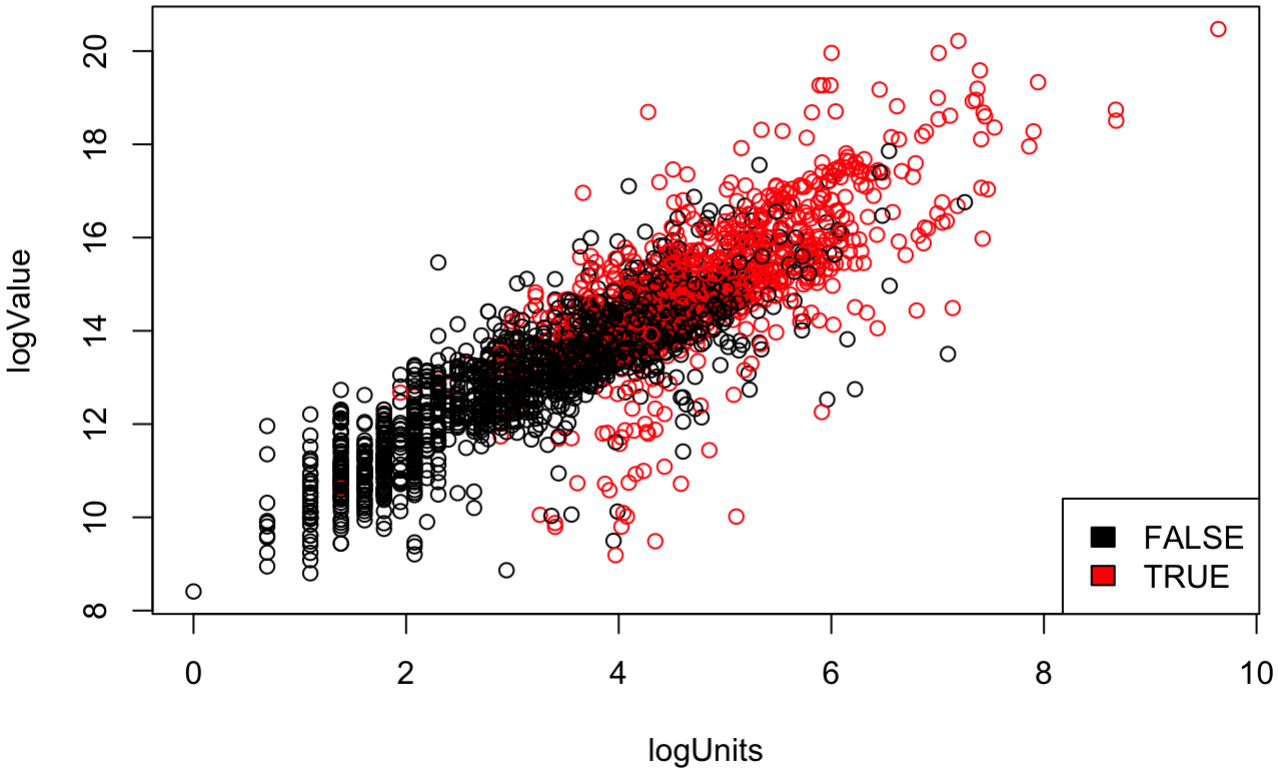
i.

```
# Plot property logValue against property logUnits
plot(housing$logUnits, housing$logValue, xlab = 'logUnits', ylab = 'logValue')
```



ii.

```
# Make a new plot adding the fator after1950
plot(housing$logUnits, housing$logValue, xlab = 'logUnits', ylab = 'logValue', col = factor(hous
ing$after1950))
legend("bottomright", legend = levels(factor(housing$after1950)), fill =unique(factor(housing$af
ter1950)))
```



From this plot, we could see that for housing built after year 1950, their logUnits and logValue is higher than the housing built before year 1950. The more units there are, the higher their values. So housing built after 1950 has more units and thus their values are higher

iii.

```
cor(housing$logUnits, housing$logValue)
```

```
## [1] 0.8727348
```

Correlation between property logValue and property logUnits from the whole data is 0.87

```
man <- housing[housing$Borough == 'Manhattan',]
cor(man$logUnits, man$logValue)
```

```
## [1] 0.8830348
```

Correlation between property logValue and property logUnits from just Manhattan is 0.88

```
bro <- housing[housing$Borough == 'Brooklyn',]
cor(bro$logUnits, bro$logValue)
```

```
## [1] 0.9102601
```

Correlation between property logValue and property logUnits from just Brooklyn is 0.91

```
aft <- housing[housing$after1950 == 'TRUE',]
cor(aft$logUnits, aft$logValue)
```

```
## [1] 0.721735
```

Correlation between property logValue and property logUnits from properties built after 1950 is 0.72

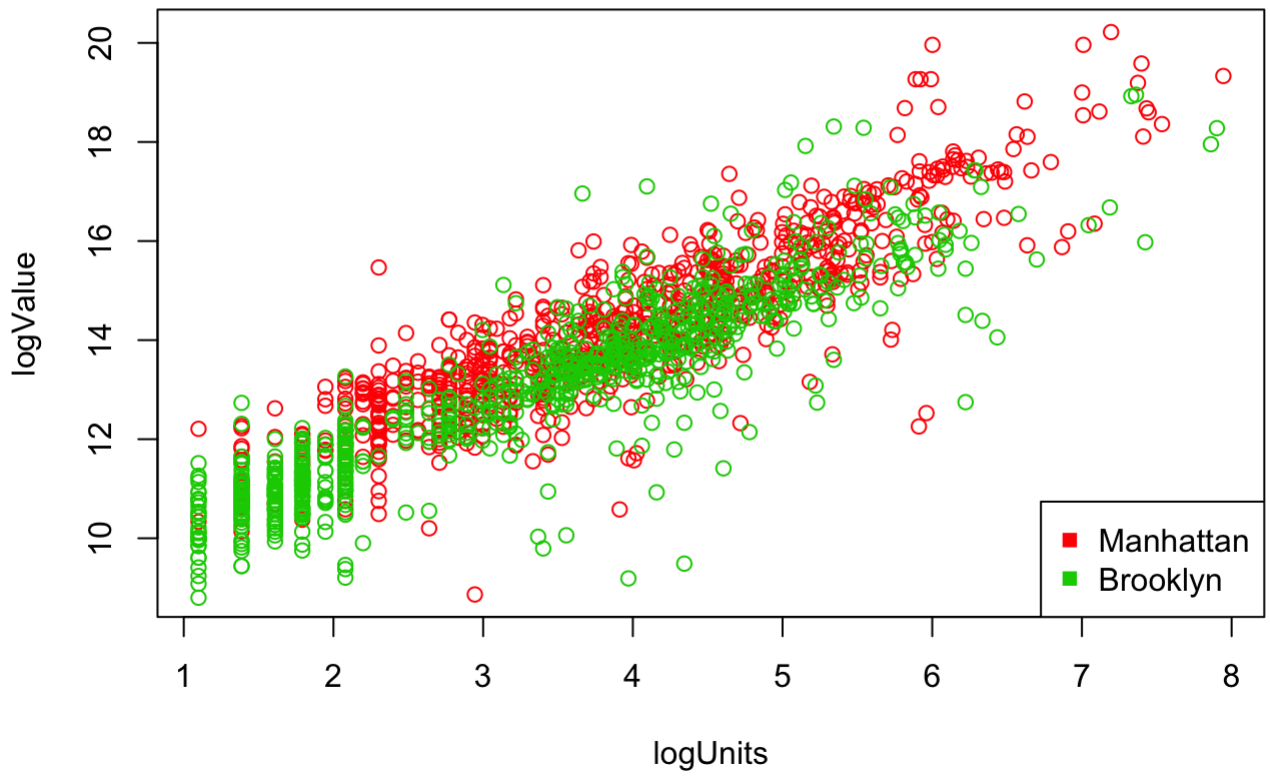
```
bef <- housing[housing$after1950 == 'FALSE',]
cor(bef$logUnits, bef$logValue)
```

```
## [1] 0.8643297
```

Correlation between property logValue and property logUnits from properties built before 1950 is 0.86

iv.

```
# Create a single plot showing property logValue against property logUnits for Manhattan and Bro
oklyn
plot(man$logUnits, man$logValue, xlab = 'logUnits', ylab = 'logValue', col = 2)
points(bro$logUnits, bro$logValue, col = 3)
legend("bottomright", legend = c("Manhattan", "Brooklyn"), col = c(2,3), pch = 15)
```



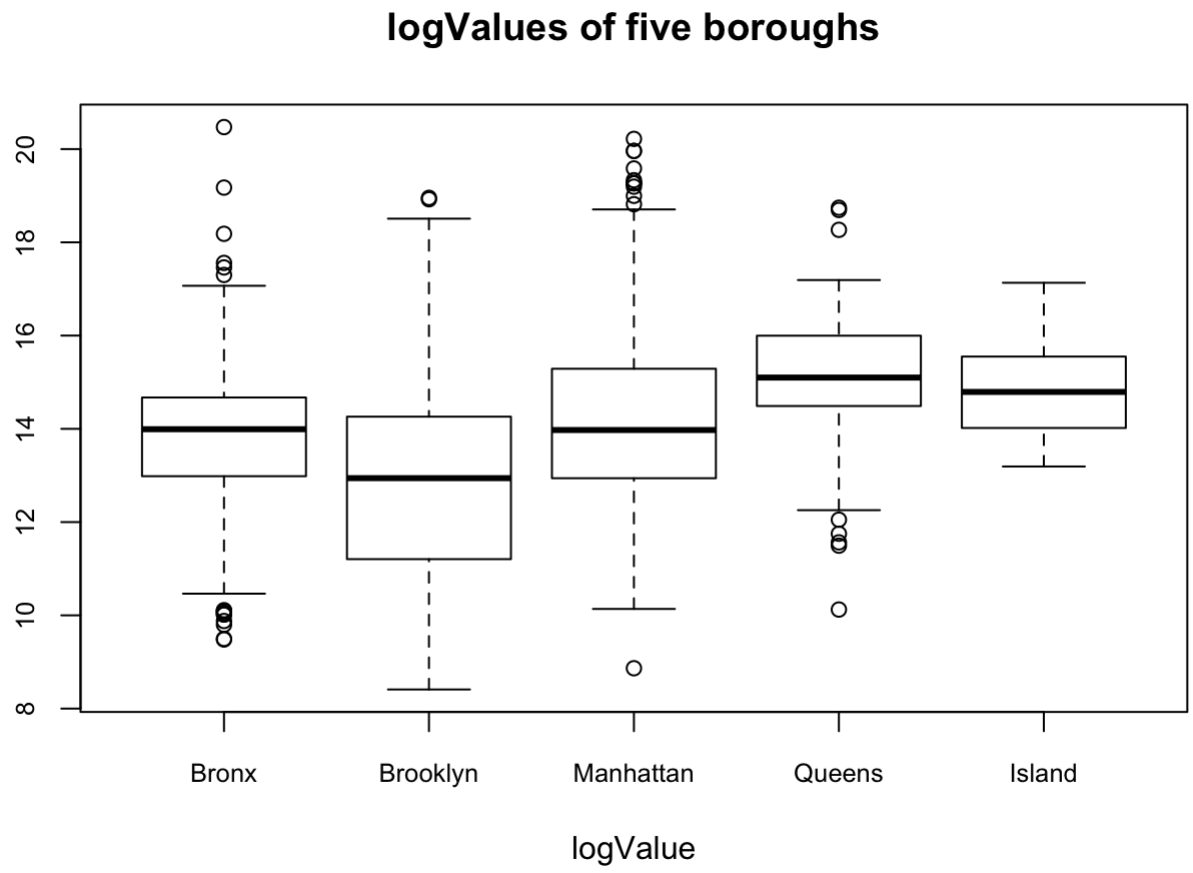
v.

```
# Use a single line to calculate the median value of Manhattan properties
median(housing$Value[housing$Borough == 'Manhattan'], na.rm = TRUE)
```

```
## [1] 1172362
```

vi.

```
# Create side by side box plots comparing property logValue across the five boroughs.
bronx <- housing$logValue[housing$Borough == 'Bronx']
brooklyn <- housing$logValue[housing$Borough == 'Brooklyn']
manhattan <- housing$logValue[housing$Borough == 'Manhattan']
queens <- housing$logValue[housing$Borough == 'Queens']
island <- housing$logValue[housing$Borough == 'Staten Island']
boxplot(bronx,brooklyn, manhattan, queens, island, names = c('Bronx', 'Brooklyn', 'Manhattan',
'Queens', 'Island'), horizontal = FALSE, main = 'logValues of five boroughs', xlab = 'logValue',
cex.axis = 0.8)
```



vii.

```
# Calculate the median property values for five boroughs.
median(housing$Value[housing$Borough == 'Bronx'])
```

```
## [1] 1192950
```

```
median(housing$Value[housing$Borough == 'Brooklyn'])
```

```
## [1] 417610
```

```
median(housing$Value[housing$Borough == 'Manhattan'])
```

```
## [1] 1172362
```

```
median(housing$Value[housing$Borough == 'Queens'])
```

```
## [1] 3611700
```

```
median(housing$Value[housing$Borough == 'Staten Island'])
```

```
## [1] 2654100
```

The median property values for Bronx is 1192950

The median property values for Brooklyn is 417610

The median property values for Manhattan is 1172362

The median property values for Queens is 3611700

The median property values for Staten Island is 2654100