Homework #1

Fan Yang (fy2232)

September 26, 2017

Part 1: Loading and Cleaning the Data in R i.

```
setwd("D:/Subject Materials/Columbia/STAT COMP & INTRO TO DATA SCI/HW")
housing <- read.csv('properties.csv')
```

ii.

```
ncol(housing)
```

```
## [1] 17
```

```
nrow(housing)
```

```
## [1] 16319
```

iii.

```
apply(is.na(housing), 2, sum)
```

```
##       cartodb_id              bbl         tract_10         sba_name
##                0                0                0                0
##         ccd_name          cd_name        boro_name        city_name
##                0                0                0                0
##  tax_delinquency    ser_violation   assessed_value       owner_name
##                0                0                0                0
##        res_units       year_built        buildings standard_address
##              504              253              319                0
##  applied_filters
##                0
```

This command gives a table of the NA values in each column.

iv.

```
vec <- which(housing$assessed_value!=0)
housing <- housing[vec,]
```

v.

```
16319-nrow(housing)
```

```
## [1] 66
```

vi.

```
logValue <- c()
logValue <- log(housing$assessed_value)
housing <- cbind(housing, logValue)
min(logValue)
```

```
## [1] 5.877736
```

```
median(logValue)
```

```
## [1] 13.2497
```

```
mean(logValue)
```

```
## [1] 13.48347
```

```
max(logValue)
```

```
## [1] 20.03494
```

vii.

```
logUnits <- log(housing$res_units)
housing <- cbind(housing, logUnits)
```

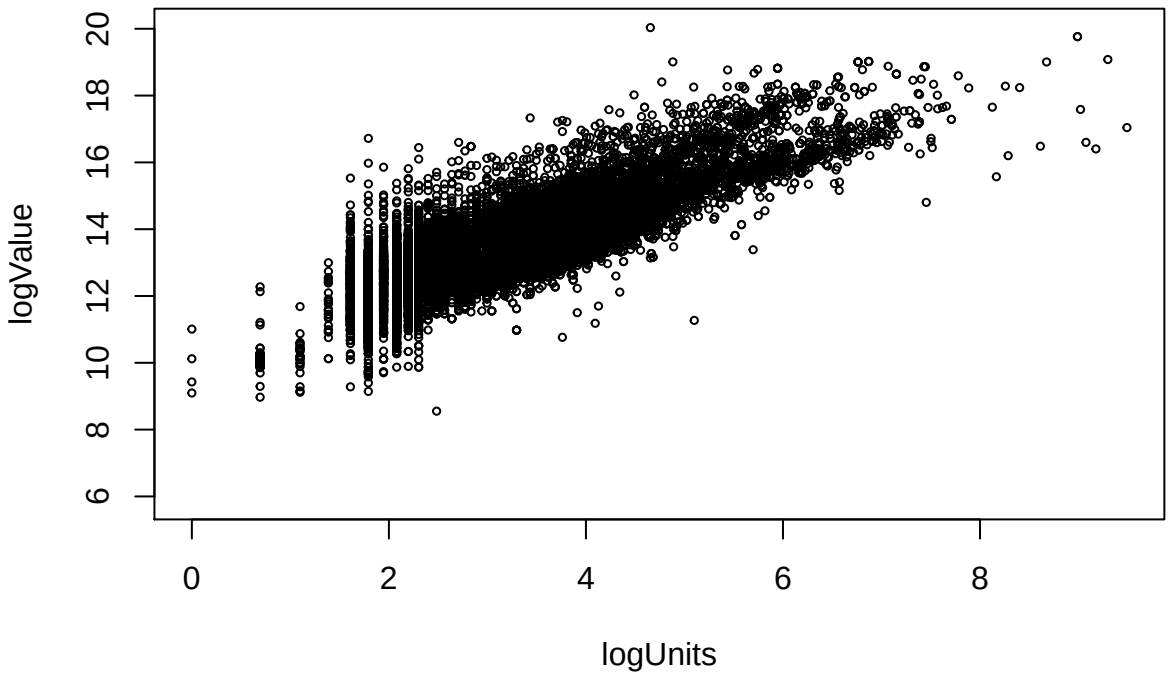
viii.

```
housing <- cbind(housing, after2000 = housing$year_built>=2000)
```

Part 2: EDA

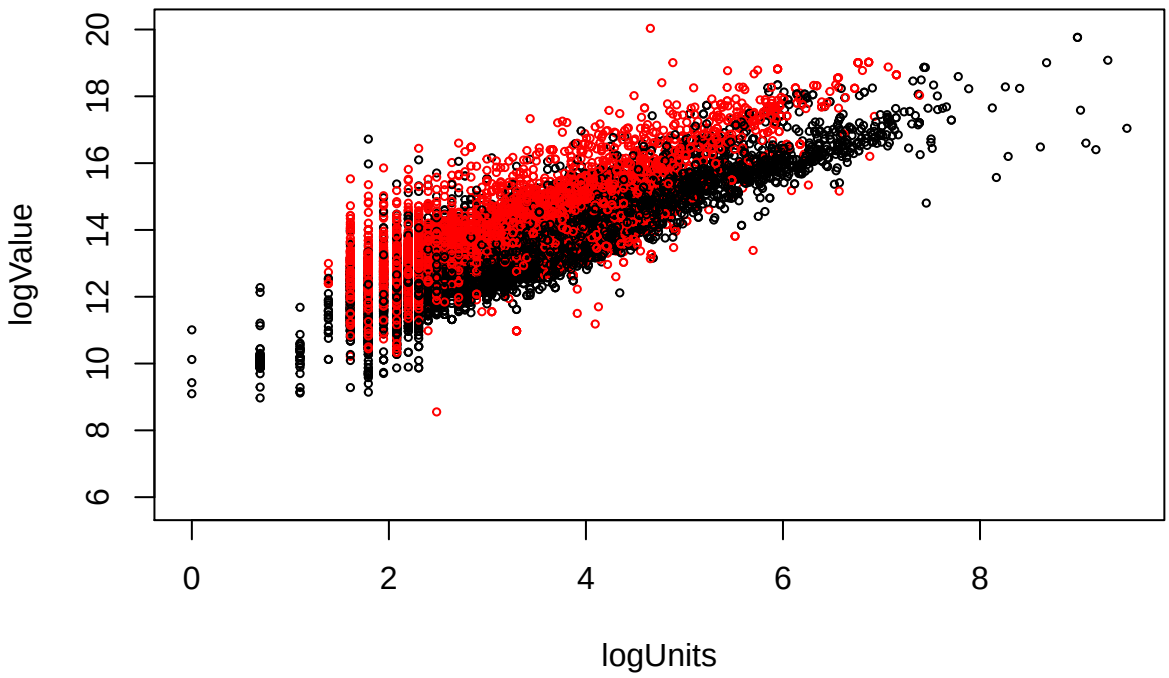
i.

```
plot(housing$logUnits,housing$logValue,xlab="logUnits",ylab="logValue",cex=0.5)
```



ii.

```
plot(housing$logUnits,housing$logValue,
     col = factor(housing$after2000),
     xlab="logUnits",ylab="logValue",cex=0.5)
```




```
cov(housing$logUnits,housing$logValue,use="pairwise.complete.obs")
```

```
## [1] 1.504415
```

The variation between the two variables is 1.504415. In the plot we can see that the red points are almost above the black points at the same x. Which tells us that the value of housing built after 2000 is greater than that before 2000. And with the units in the property increase, the value also increase.

iii. (i)the whole data,

```
cor(housing$logUnits,housing$logValue,use="pairwise.complete.obs")
```

```
## [1] 0.8431877
```

(ii) just Manhattan

```
cor(housing$logUnits[housing$boro_name=='Manhattan'],
    housing$logValue[housing$boro_name=='Manhattan'],
    use="pairwise.complete.obs")
```

```
## [1] 0.8592745
```

(iii) just Brooklyn

```
cor(housing$logUnits[housing$boro_name=='Brooklyn'],
    housing$logValue[housing$boro_name=='Brooklyn'],
    use="pairwise.complete.obs")
```

```
## [1] 0.8579328
```

(iv) for properties built after 2000

```r
cor(housing$logUnits[housing$after2000],
    housing$logValue[housing$after2000],
    use="pairwise.complete.obs")
```

```
## [1] 0.8337845
```

(v) for properties built before 2000

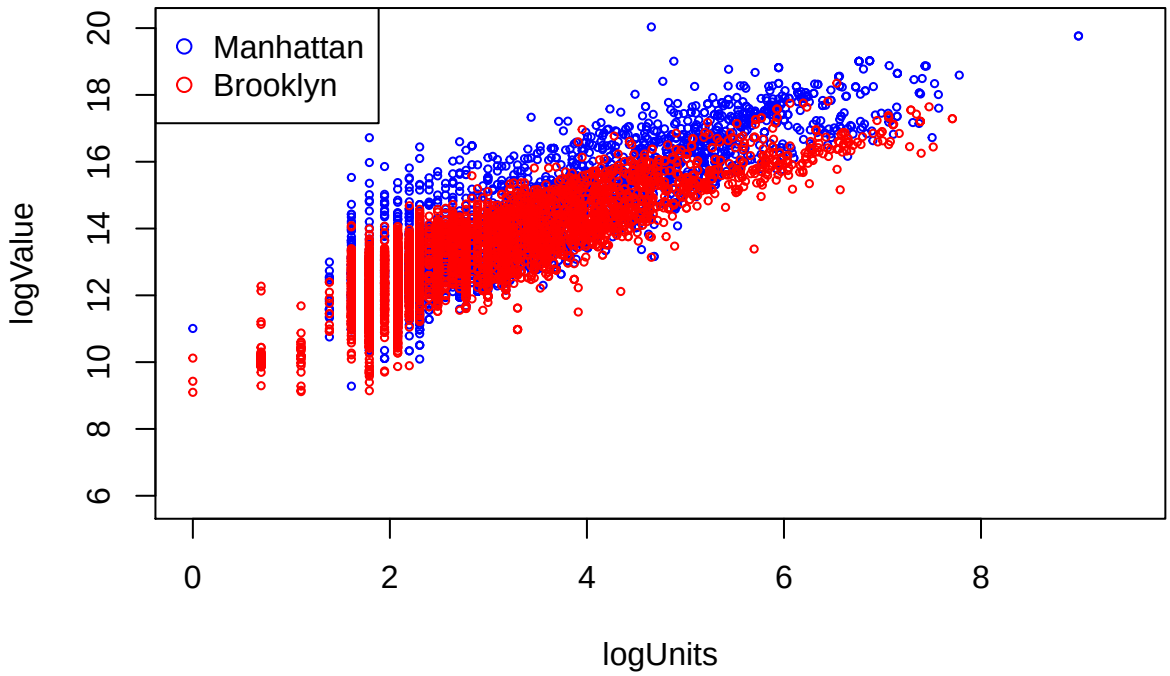
```r
cor(housing$logUnits[housing$year_built<2000],
    housing$logValue[housing$year_built<2000],
    use="pairwise.complete.obs")
```

```
## [1] 0.8927153
```

iv.

```r
plot(range(housing$logUnits, na.rm = TRUE),
     range(housing$logValue, na.rm = TRUE),
     xlab= "logUnits", ylab = "logValue",type='n')

lines(housing$logUnits[housing$boro_name=='Manhattan'],
      housing$logValue[housing$boro_name=='Manhattan'],
      type='p',col="blue",cex=0.5)
lines(housing$logUnits[housing$boro_name=='Brooklyn'],
      housing$logValue[housing$boro_name=='Brooklyn'],
      type='p',col="red",cex=0.5)
legend("topleft",c("Manhattan","Brooklyn"),col=c("blue","red"),pch=1)
```

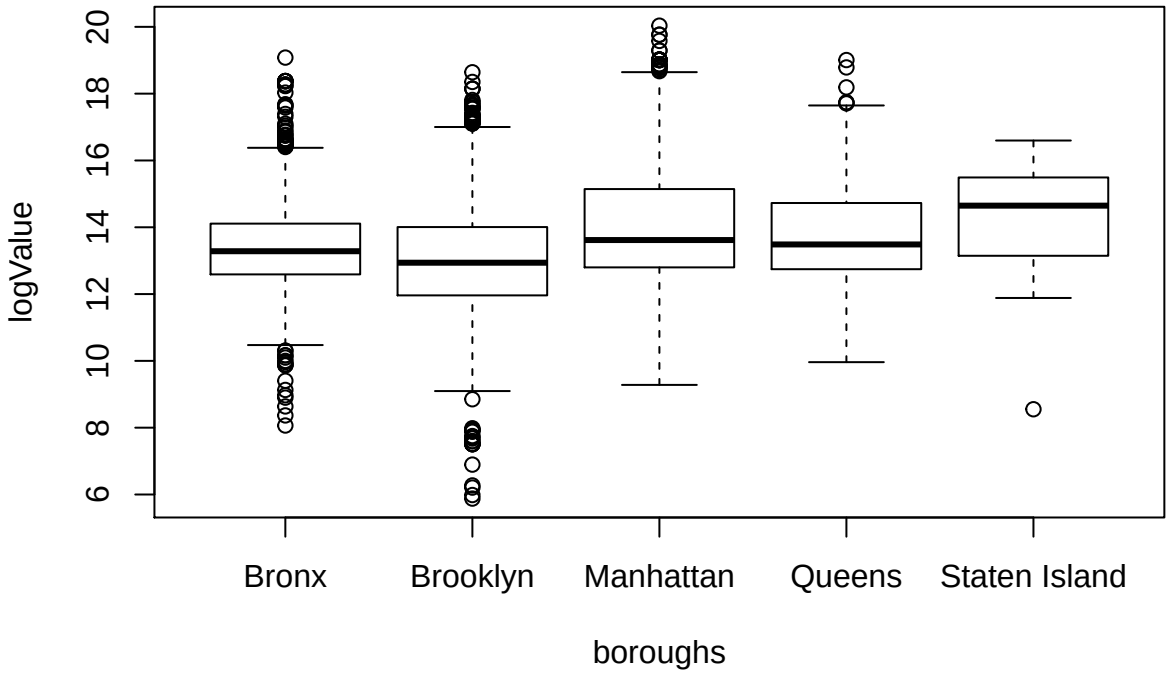


v.

```
med.value <- median(housing$assessed_value[housing$boro_name=='Manhattan'], na.rm = TRUE)
```

vi.

```
boxplot(housing$logValue ~ housing$boro_name,
        ylab = "logValue", xlab = "boroughs")
```



vi.

```
tapply(housing$assessed_value,housing$boro_name,median)
```

```
##         Bronx      Brooklyn     Manhattan        Queens Staten Island
##        587250        416014        820350        719100       2296350
```