hw2

NAME: Yuhao Wang, UNI: yw3204

9/22/2018

Part1

i.

```
# load data
housing <- read.csv("NYChousing.csv")
```

ii.

```
# no. of rows and colums
dim(housing)
```

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

iii.

```
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
```

It is finding the number of all “NA”" records for each variable or column.

iv.

```r
# remove the rows for which the variable Value is NA.
ind <- is.na(housing$Value)
ind = !ind
housing <- housing[ind, ]
```

v.

It removes 52 rows and agrees with the results from (iii).

vi.

```r
# create logarithm of "Value"
housing$logValue <- log(housing$Value, 10)
summary(housing$logValue)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.652   5.425   5.971   5.942   6.429   8.891
```

The minimum, median, mean, and maximum values of logValue are listed above in the summary.

vii.

```r
# create logarithm of "UnitCount"
housing$logUnits <- log(housing$UnitCount, 10)
```

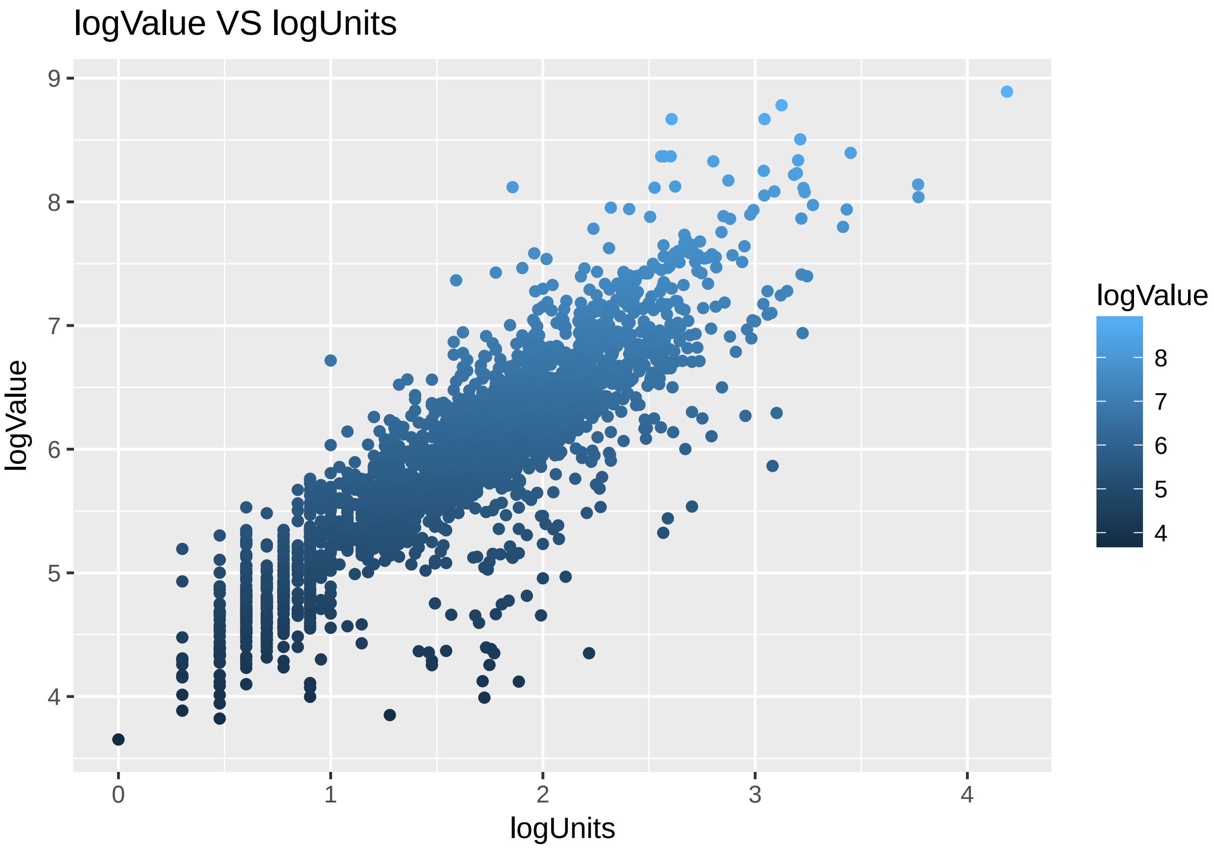
viii.

```r
# create a new variable "after1950"
housing$after1950 <- ifelse(housing$YearBuilt >= 1950, TRUE, FALSE)
```

Part2

i.

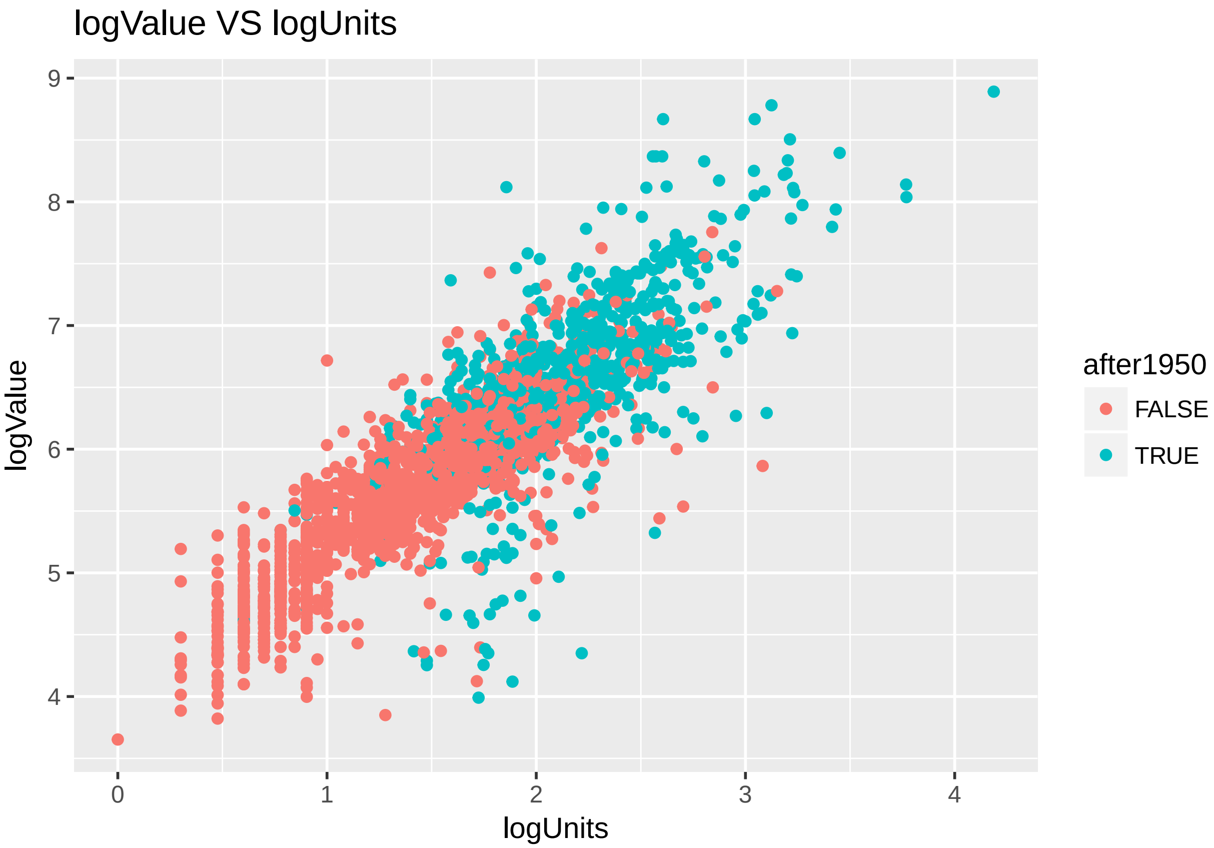
```r
# plot logValue against logUnits
library(ggplot2)
ggplot(housing, aes(x = logUnits, y = logValue)) + geom_point(aes(color = logValue)) +
  labs(x = "logUnits", y = "logValue", title = "logValue VS logUnits")
```



ii.

```
# color data by "after1950"

# solution 1
ggplot(housing, aes(x = logUnits, y = logValue)) + geom_point(aes(color = after1950)) +
  labs(x = "logUnits", y = "logValue", title = "logValue VS logUnits")
```



```
# solution 2
# plot(housing$logUnits, housing$logValue, col = factor(housing$after1950),
#   xlab = "logUnits", ylab = "logValue")
# legend("bottomright", legend = levels(factor(housing$after1950)),
#   fill = unique(factor(housing$after1950)))
```

Basically, the larger the logUnits, the larger the logValue. And in general, logUnits and logValue after 1950 tend to be larger than that before 1950.

iii.

```
# calculate correlation
cor(housing[, c("logValue", "logUnits")])
```

```
##           logValue  logUnits
## logValue 1.0000000 0.8727348
## logUnits 0.8727348 1.0000000
```

```
ind1 <- housing$Borough == "Manhattan"
cor(housing[ind1, c("logValue", "logUnits")])
```

```
##           logValue  logUnits
## logValue 1.0000000 0.8830348
## logUnits 0.8830348 1.0000000
```

```
ind2 <- housing$Borough == "Brooklyn"
cor(housing[ind2, c("logValue", "logUnits")])
```

```
##            logValue  logUnits
## logValue 1.0000000 0.9102601
## logUnits 0.9102601 1.0000000
```

```
ind3 <- housing$after1950 == TRUE
cor(housing[ind3, c("logValue", "logUnits")])
```

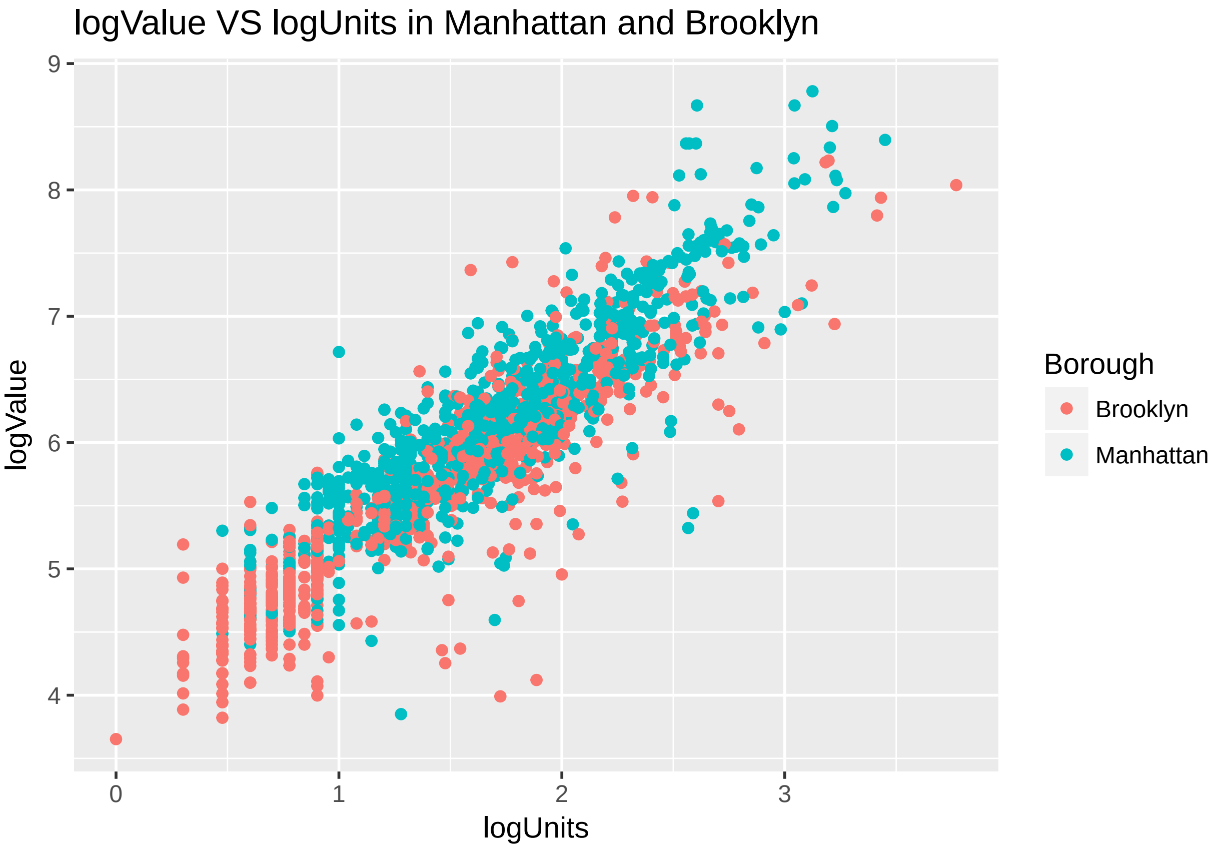
```
##           logValue logUnits
## logValue 1.000000 0.721735
## logUnits 0.721735 1.000000
```

```
ind4 <- housing$after1950 == FALSE
cor(housing[ind4, c("logValue", "logUnits")])
```

```
##            logValue  logUnits
## logValue 1.0000000 0.8643297
## logUnits 0.8643297 1.0000000
```

iv.

```
# plot logValue against logUnits for Manhattan and Brooklyn
ind5 <- housing$Borough == "Manhattan" | housing$Borough == "Brooklyn"
sub <- housing[ind5, c("logValue", "logUnits", "Borough")]
ggplot(sub, aes(x = logUnits, y = logValue)) + geom_point(aes(color = Borough)) +
  labs(x = "logUnits", y = "logValue", title = "logValue VS logUnits in Manhattan and Brooklyn")
```



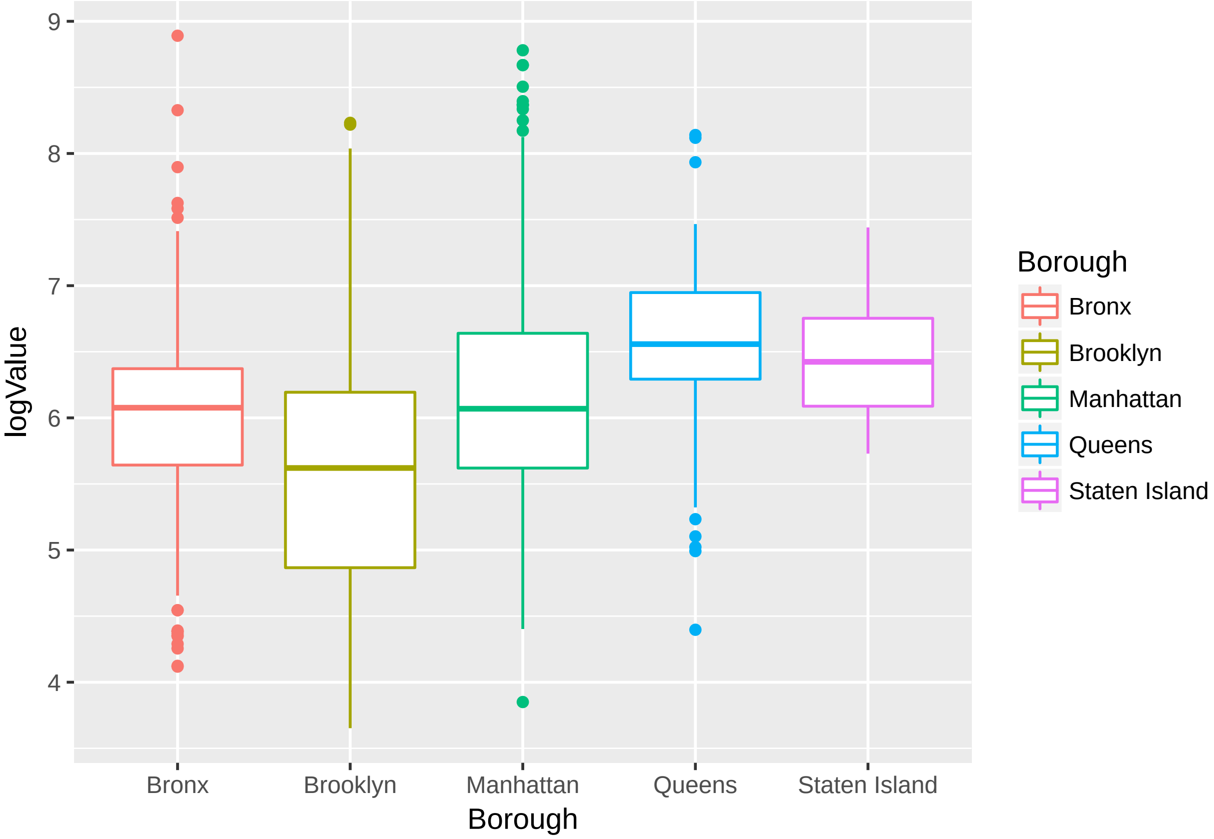
v.

```
# find the median of "Value" in Manhattan
median(housing[housing$Borough == "Manhattan", ]$Value, na.rm = TRUE)
```

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

vi.

```
# box plots comparing logValue across the five boroughs.
ggplot(housing, aes(x = Borough, y = logValue)) + geom_boxplot(aes(color = Borough))
```



vii.

```
# calculate median for Value for five boroughs
temp <- c()
for(name in c("Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island")) {
  temp <- c(temp, median(housing[housing$Borough == name, ]$Value, na.rm = TRUE))
}
temp
```

```
## [1] 1192950  417610 1172362 3611700 2654100
```

As above, the median of Value for Bronx, Brooklyn, Manhattan, Queens and Staten Island are 1192950, 417610, 1172362, 3611700 and 2654100.