Lab 5

Salim M'jahad msm2243

February 15, 2018

Part 1

i.

```
housing = read.csv("NYChousing.csv")
```

ii.

```
dim(housing)
```

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

The dataframe has 2506 rows and 22 columns

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
```

This function applies the function sum on the values in housing that are NA and returns the results in a vector format.

iv.

```
housing <- subset(housing, !is.na(Value))
```

v.

```
dim(housing)
```

```
## [1] 2454   22
```

2454 = 2506 - 52

I removed 52 rows, so yes it agrees.

vi.

```
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
```

minimum = 8.41 median = 13.75 mean = 13.68 maximum = 20.47

vii.

```
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.

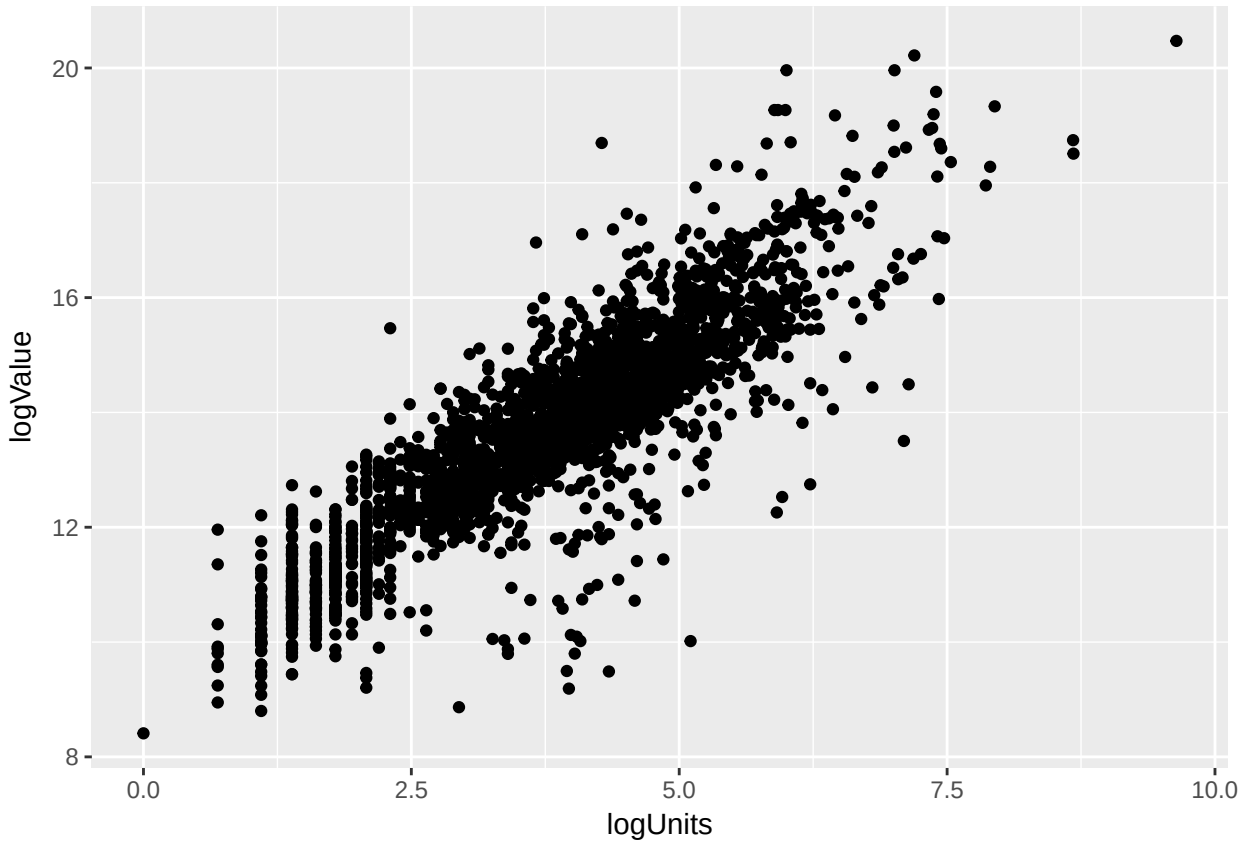
```
housing$after1950 <- housing$YearBuilt >= 1950
summary(housing$after1950)
```

```
##    Mode   FALSE    TRUE
## logical    1594     860
```

Part 2

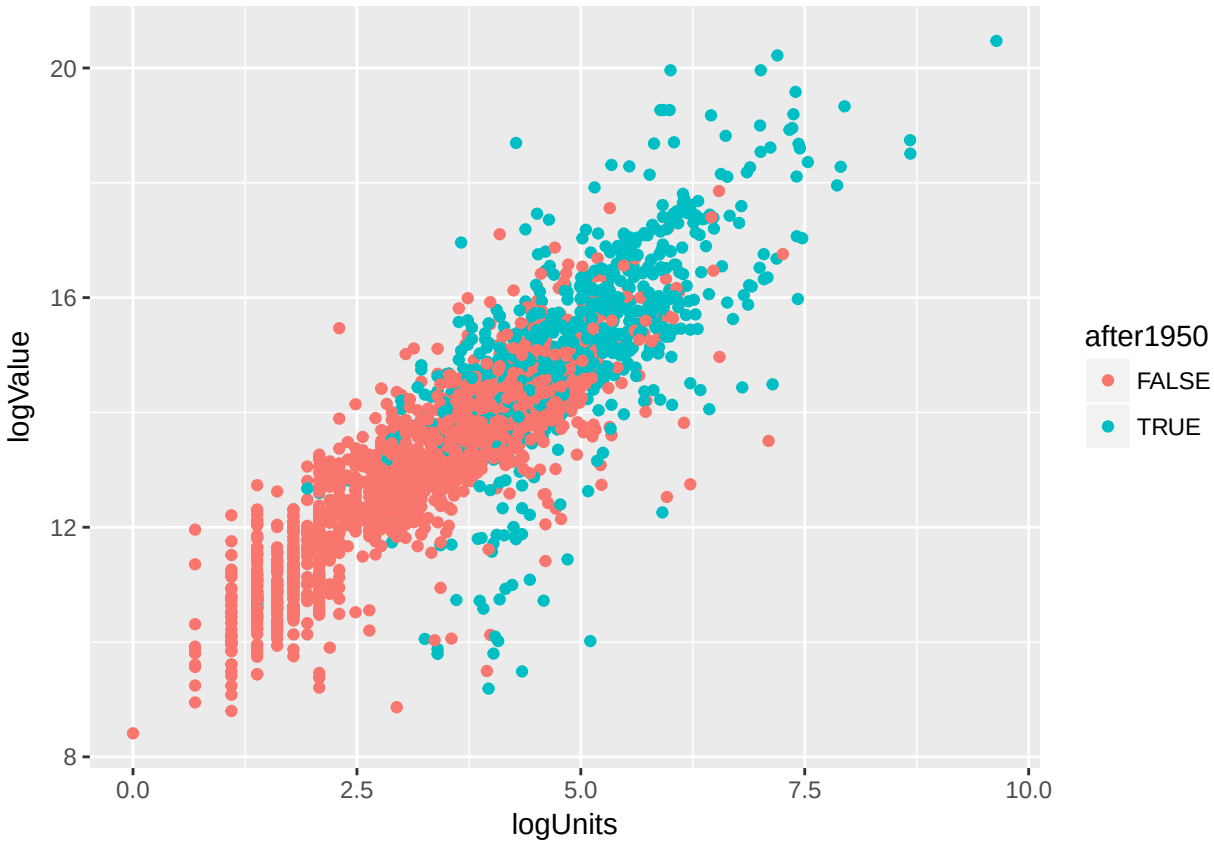
i.

```
library(ggplot2)

p1 <- ggplot(data = housing) + geom_point(mapping = aes(y = logValue, x = logUnits))
p1
```



ii.

```
p2 <- ggplot(data = housing) + geom_point(mapping = aes(y = logValue, x = logUnits, col=after1950))
p2
```



There is almost a direct linear relationship between the logUnits variable and the logValue variable. The more units there are in a property the larger its value. The covariance seems to be positive.

iii.

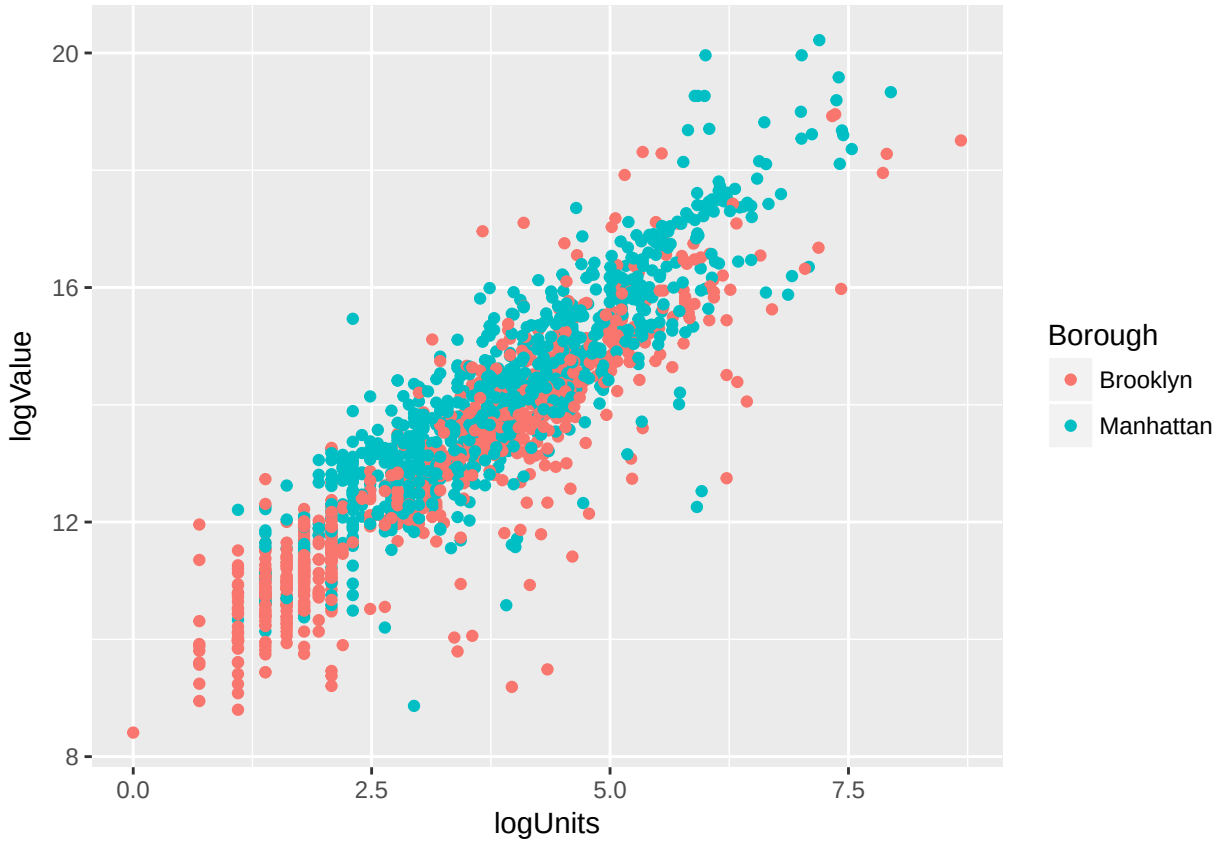
```
man <- subset(housing, Borough=="Manhattan")
brook <- subset(housing, Borough=="Brooklyn")
a1950 <- subset(housing, after1950 == TRUE)
b1950 <- subset(housing, after1950 == FALSE)

cov_all <- cov(x=housing$logValue, y=housing$logUnits)
cov_man <- cov(x=man$logValue, y=man$logUnits)
cov_brook <- cov(x=brook$logValue, y=brook$logUnits)
cov_a1950 <- cov(x=a1950$logValue, y=a1950$logUnits)
cov_b1950 <- cov(x=b1950$logValue, y=b1950$logUnits)

c(cov_all, cov_man, cov_brook, cov_a1950, cov_b1950)
```

```
## [1] 2.182148 1.983556 2.566733 1.123973 1.519853
```

(i) the whole data: 2.182148
(ii) just Manhattan: 1.983556
(iii) just Brooklyn: 2.566733
(iv) for properties built after 1950: 1.123973
(v) for properties built before 1950: 1.519853

iv.

```
manorbrook <- subset(housing, Borough %in% c("Manhattan", "Brooklyn"))
p3 <- ggplot(data = manorbrook) + geom_point(mapping = aes(y = logValue, x = logUnits, col =Borough))
p3
```



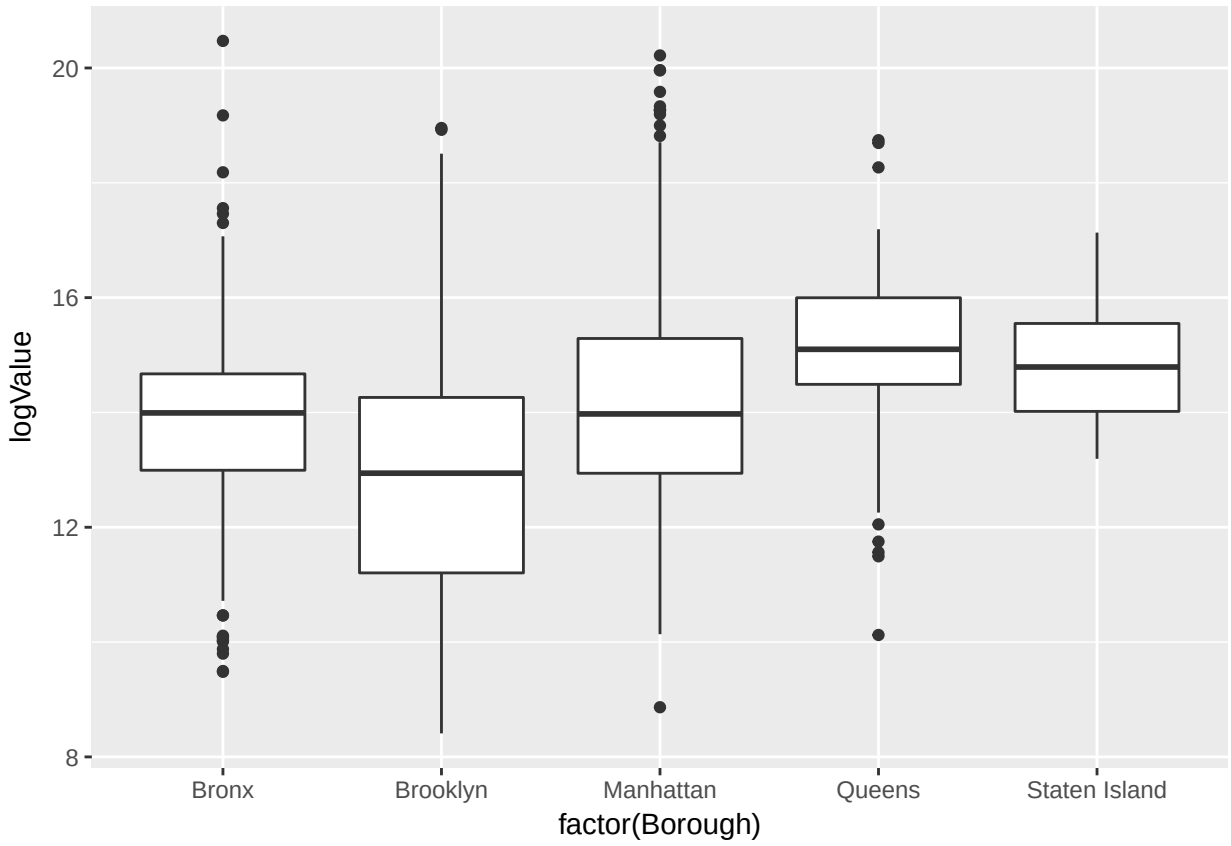
v.

```
median(subset(housing, Borough=="Manhattan")$Value, na.rm = TRUE)
```

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

vi.

```
p4 <- ggplot(data=housing) + geom_boxplot(aes(x=factor(Borough), y=logValue))
p4
```



vii.

```
X <- split(housing, housing$Borough)

medi <- function(datf) {
  return(median(datf$Value, na.rm = TRUE))
}
sapply(X, medi)
```

```
##         Bronx      Brooklyn     Manhattan        Queens Staten Island
##       1192950        417610       1172362       3611700       2654100
```