HW2_mz2692 GR5206

Steven

9/25/2018

Part 1: Loading and Cleaning the Data in R

i. Load the data into a dataframe calledhousing.

```
housing=as.data.frame(read.csv("/Users/zhongming/Downloads/NYChousing.csv"))
```

ii. How many rows and columns does the dataframe have?

```
dim(housing)
```

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

There are 2506 rows and 22 columns.

iii. Run this command, and explain, in words, what this does:

```
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 code calculates the number of NAs in different columns.

iv. Remove the rows of the dataset for which the variableValueis NA.

```
housing=na.omit(housing)
```

v. How many rows did you remove with the previous call? Does this agree with your resultfrom (iii)?

```
2506-dim(housing)[1]
```

```
## [1] 1876
```

1876 rows have been removed. This agrees with the reuslt in (iii).

vi. Create a new variable in the dataset calledlogValuethat is equal to the logarithm ofthe property'sValue. What are the minimum, median, mean, and maximum values oflogValue?

```
housing["logValue"]=log(housing["Value"])
summary(housing["logValue"])
```

```
##     logValue
##  Min.   :10.06
##  1st Qu.:13.82
##  Median :14.65
##  Mean   :14.65
##  3rd Qu.:15.38
##  Max.   :20.22
```

The minimum is 10.06, the median is 14.65, the mean is 14.65, the maximum is 20.22.

vii. Create a new variable in the dataset calledlogUnitsthat is equal to the logarithm ofthe number of units in the property. The number of units in each piece of property isstored in the variableUnitCount.

```
housing["logUnits"]=log(housing["UnitCount"])
```

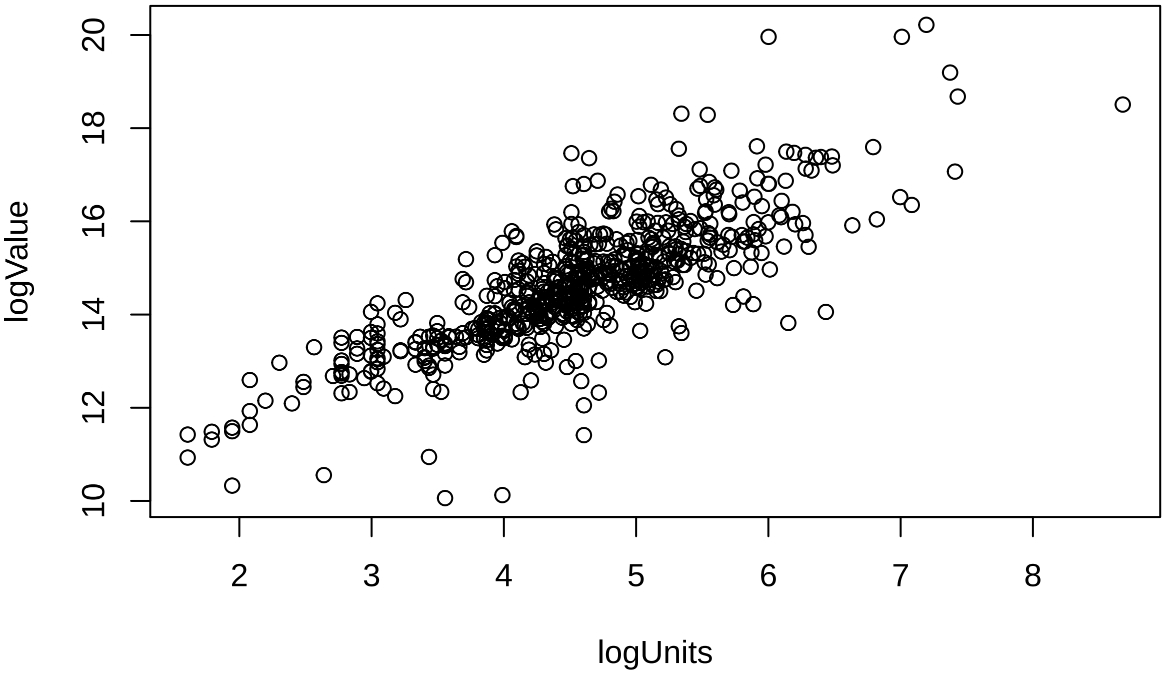
viii. Finally create a new variable in the dataset calledafter1950which equalsTRUEifthe property was built in or after 1950 andFALSEotherwise. You'll want to use theYearBuiltvariable here. This can be done in a single line of code.

```
housing["after1950"]=housing["YearBuilt"]>=1950
```

Part 2: EDA

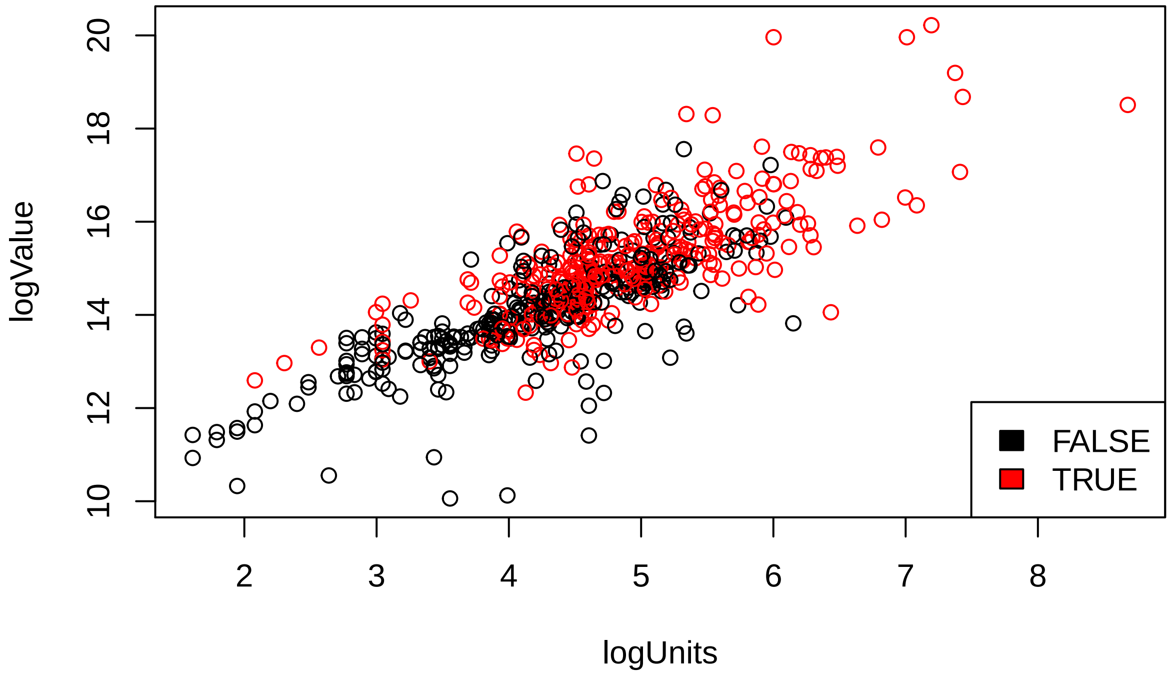
i. Plot propertylogValueagainst propertylogUnits. Name the x and y labels of theplot appropriately.logValueshould be on the y-axis.

```
with(housing,plot(x=logUnits,y=logValue,xlab="logUnits",ylab="logValue"))
```



ii. Make the same plot as above, but now include the argumentcol = factor(housing$after1950).Describe this plot and the covariation between the two variables. What does the coloringin the plot tell us?

```
with(housing,plot(x=logUnits,xlab="logUnits",y=logValue,ylab="logValue",col= factor(housing$after1950))
legend("bottomright",legend=levels(factor(housing$after1950)),fill= unique(factor(housing$after1950)))
```



The plot can be approximate to a liner relationship, which shows that the covariance between logValue and log Units is positive. The coloring in the plot tell us that data of buidlings builted in or after 1950 are in read dots and data of buildiings builted before 1950 are in black dots.

iii. Thecor()function calculates the correlation coefficient between two variables. Whatis the correlation between propertylogValueand propertylogUnitsin (i) the wholedata, (ii) just Manhattan (iii) just Brooklyn (iv) for properties built after 1950 (v) forproperties built before 1950?

```
#(i) the wholedata
cor(data.frame(x=housing$logUnits,y=housing$logValue))
```

```
##           x         y
## x 1.0000000 0.7988655
## y 0.7988655 1.0000000
```

```
#(ii) just Manhattan
cor(data.frame(x=housing[housing$Borough=="Manhattan","logValue"],y=housing[housing$Borough=="Manhattan
```

```
##           x         y
## x 1.0000000 0.8710823
## y 0.8710823 1.0000000
```

```
#(iii) just Brooklyn
cor(data.frame(x=housing[housing$Borough=="Brooklyn","logValue"],y=housing[housing$Borough=="Brooklyn",
```

```
##           x         y
## x 1.0000000 0.8053241
## y 0.8053241 1.0000000
```

```
#(iv) for properties built after 1950
cor(data.frame(x=housing[housing$after1950==TRUE,"logValue"],y=housing[housing$after1950==TRUE,"logUnit
##          x        y
## x 1.000000 0.746731
## y 0.746731 1.000000
```

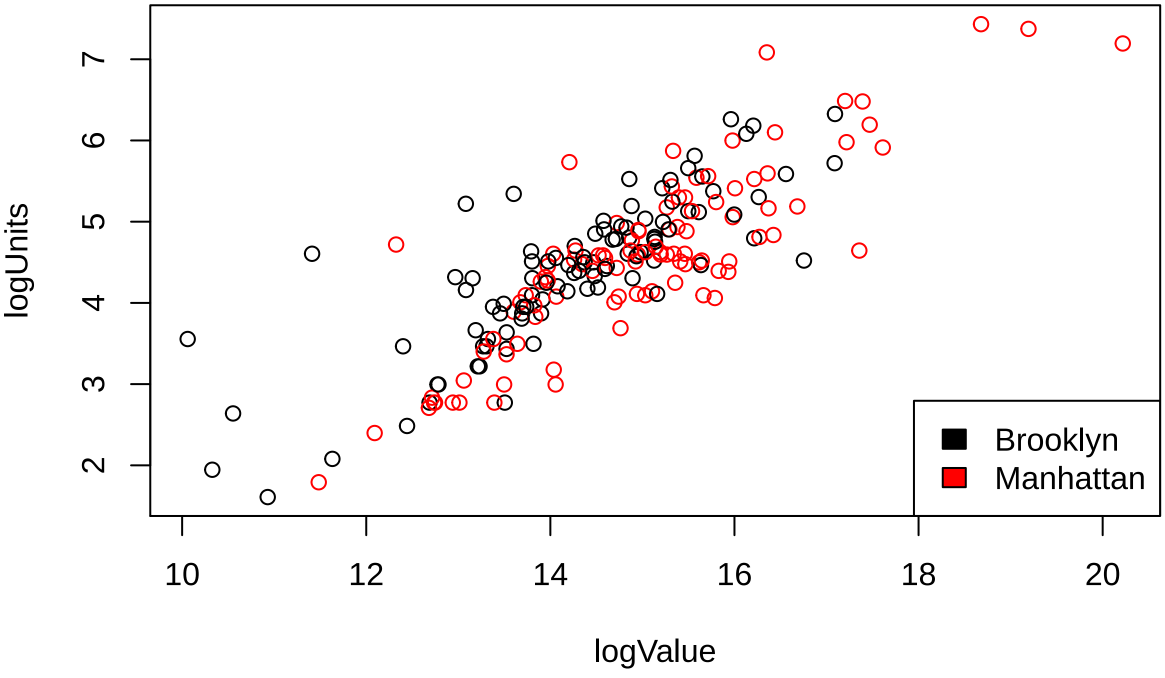
```
#(v) forproperties built before 1950
cor(data.frame(x=housing[housing$after1950==F,"logValue"],y=housing[housing$after1950==F,"logUnits"]))
```

```
##           x         y
## x 1.0000000 0.7720285
## y 0.7720285 1.0000000
```

the correlation between propertylogValueand propertylogUnitsin (i) the wholedata=0.7988655 (ii) just Manhattan=0.8710823 (iii) just Brooklyn=0.8053241 (iv) for properties built after 1950=0.746731 (v) forproperties built before 1950=0.7720285

iv. Make a single plot showing propertylogValueagainst propertylogUnitsfor Manhat-tan and Brooklyn. When creating this plot, clearly distinguish the two boroughs.

```
df=data.frame(housing[housing$Borough==c("Brooklyn","Manhattan"),c("Borough","logValue","logUnits")])
plot(df[-1],col=factor(df$Borough))
legend("bottomright",legend=levels(factor(df$Borough)),fill= unique(factor(df$Borough)))
```



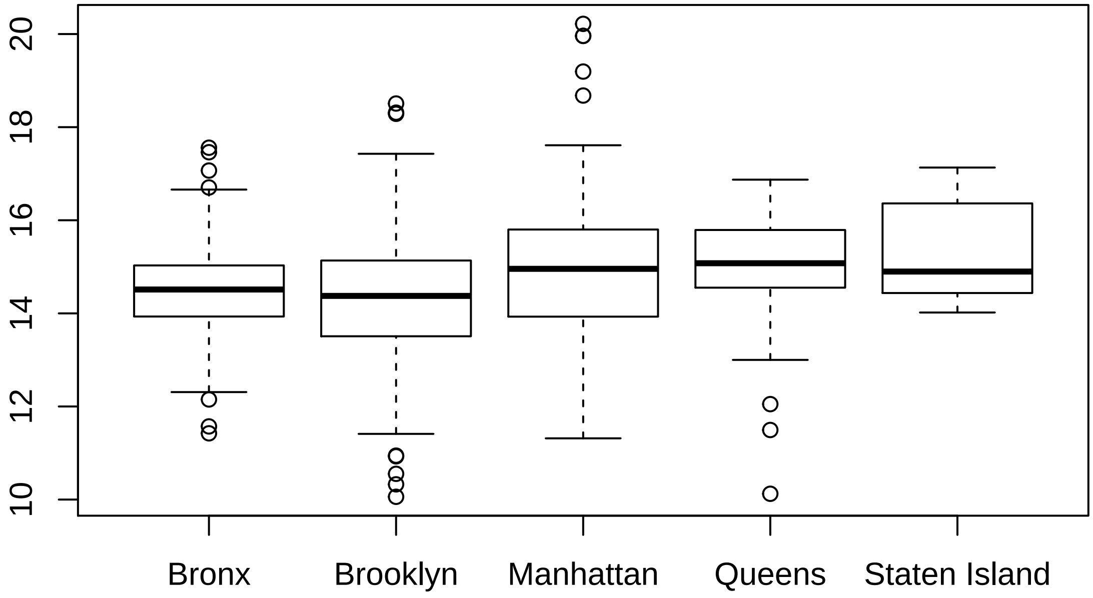
v. Consider the following block of code. Give a single line of R code which gives the samefinal answer as the block of code. There are a few ways to do this.

```
median(housing[housing$Borough=="Manhattan","Value"],na.rm=T)
```

```
## [1] 3129300
```

vi. Make side-by-side box plots comparing propertylogValueacross the five boroughs.

```r
boxplot(housing$logValue~housing$Borough)
```



vii. For five boroughs, what are the median property values? (UseValuehere, notlogValue.)

```r
aggregate(housing$Value,list(housing$Borough),median)
```

```
##         Group.1       x
## 1         Bronx 2008260
## 2      Brooklyn 1749465
## 3     Manhattan 3129300
## 4        Queens 3529800
## 5 Staten Island 2952900
```