GR5206 - HW1

Mathieu Sauterey - UNI: mjs2364

September 22, 2017

Part 1: Loading and Cleaning the Data in R

i. Load the data into a dataframe called housing

```
housing <- read.csv("properties.csv",sep="," , header = TRUE, as.is =TRUE) #Import data
```

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

```
print(dim1 <- dim(housing)) #Display dataframe dimensions
```

```
## [1] 16319    17
```

The dataframe has 16,319 rows of property data (and 1 row of header that we used in read.csv function to name the dataframe columns) and 17 columns.

iiii. Run apply(is.na(housing), 2, sum) and explain, in words, what it does

It sums up the number of instances where no information is available (NA) in each column of the dataframe

iv. Remove the rows of the dataset for which the variable assessed_value equals 0

```
housing <- housing[housing$assessed_value!=0,] #Filters out rows of dataset with "0" value
```

v. How many rows did you remove with the previous call?

```
dim2<- dim(housing) #computes new dimensions
print(dim1[1]-dim2[1]) #prints the number of rows with "0" value that were deleted
```

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

66 rows were deleted following the previous call.

vi. Create a new variable in the dataset called logValue that is equal to the logarithm of the property's assessed value. What are the minimum, median, mean, and maximum values of logValue?

```
logValue<-log(housing$assessed_value) #calculates log of properties values
min(logValue) #prints the minimum of logValues
```

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

```
max(logValue) #prints the maximum of logValues
```

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

```
median(logValue) #prints the median of logValues
```

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

```
mean(logValue) #prints the mean of logValues
```

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

The minimum of logValue is 5.88, the maximum is 20.04, the median is 13.25 and the mean is 13.48.

vii. Create a new variable in the dataset called logUnits that is equal to the logarithm of the number of units in the property.

```
logUnits <- log(housing$res_units) #calculates log of number of units in properties
```

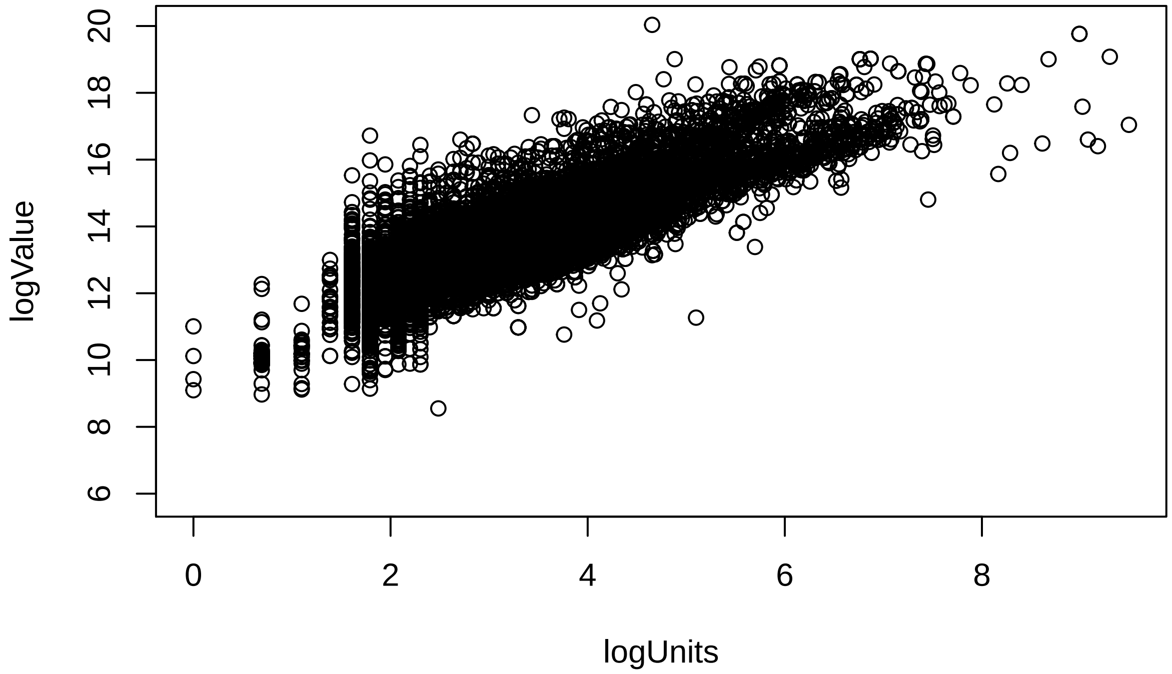
viii. Finally create a new variable in the dataset called after2000 which equals TRUE if the property was built in or after 2000 and FALSE otherwise.

```
housing$after2000 <- (housing$year_built>=2000)
#creates new column of logical type which indicates if properties where built in or after 2000
```

Part 2: EDA

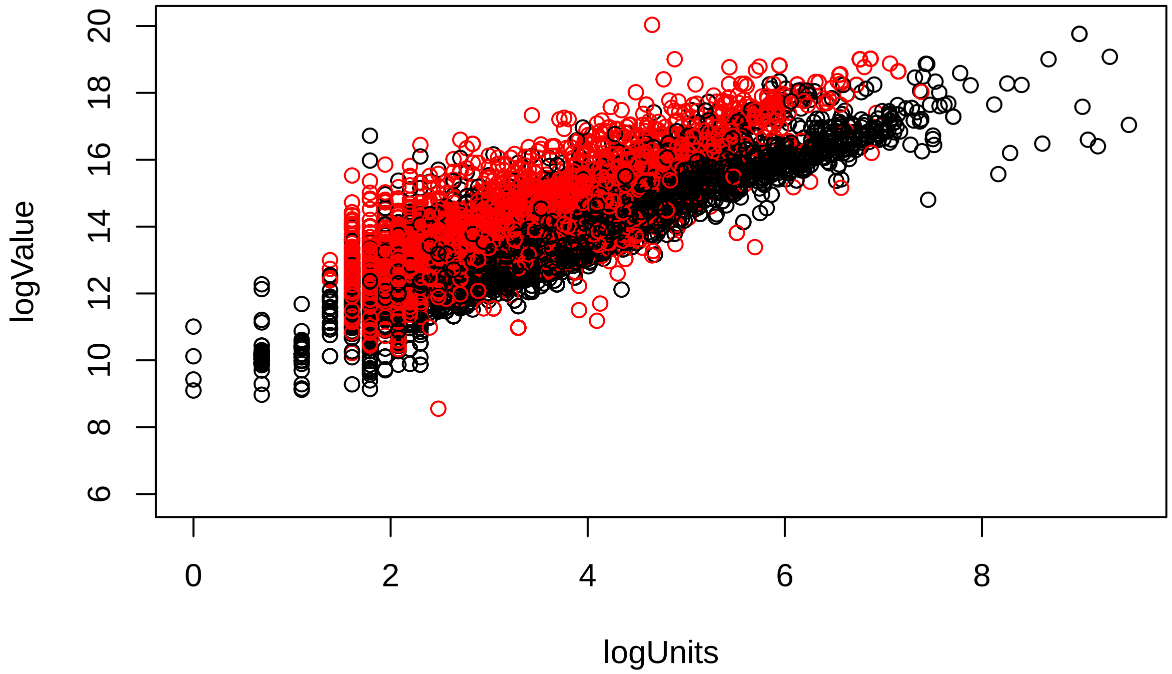
i. Plot property logValue against property logUnits. Name the x and y labels of the plot appropriately. logValue should be on the y-axis.

```
plot(logUnits,logValue,xlab="logUnits",ylab="logValue") #plots logValue vs logUnits
```



ii. Make the same plot as above, but now include the argument col=factor(housing$after2000) Describe this plot and the covariation between the two variables. What does the coloring in the plot tell us?

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



```
#plots the sme data but highlights those properties built after 2000
```

There is a linear relation with positive slope between the logUnits and the logValue variables. The more units in a property, the larger the value. Also, the coloring tell us that the majority of properties built in or after 2000 have larger value than the older properties.

iii. What is the correlation between property logValue and property logUnits in (i) the whole data,(ii) just Manhattan (iii) just Brooklyn (iv) for properties built after 2000 (v) for properties built before 2000?

```
manhat <- housing$boro_name=="Manhattan" #filters so to keep only properties in Manhattan
brook <- housing$boro_name=="Brooklyn" #filters so to keep only properties in Brooklyn
modern <- housing$after2000 #filters so to keep only properties built in or after 2000
old <- housing$after2000==FALSE #filters so to keep only properties built before 2000

#Below computes the correlation coefficient between logValue and logUnits
#for properties in whole data set, just in Manhattan, just in Brooklyn,
#built in or after 2000, and before 2000.
cor(logValue,logUnits,use="pairwise.complete.obs")
```

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

```
cor(housing$assessed_value[manhat],log(housing$res_units[manhat]),use="pairwise.complete.obs")
```

```
## [1] 0.5763046
```

```
cor(housing$assessed_value[brook],log(housing$res_units[brook]),use="pairwise.complete.obs")
```

```
## [1] 0.6656859
```

```
cor(housing$assessed_value[modern],log(housing$res_units[modern]),use="pairwise.complete.obs")
```

```
## [1] 0.4779809
```

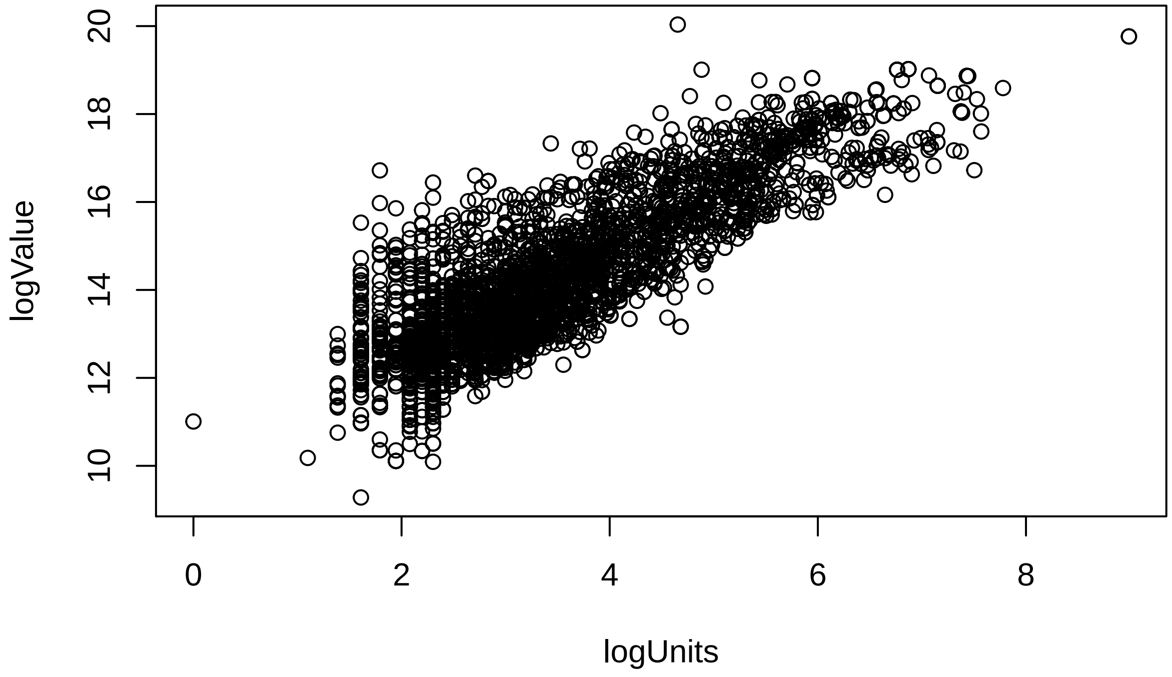
```
cor(housing$assessed_value[old],log(housing$res_units[old]),use="pairwise.complete.obs")
```

```
## [1] 0.4771511
```

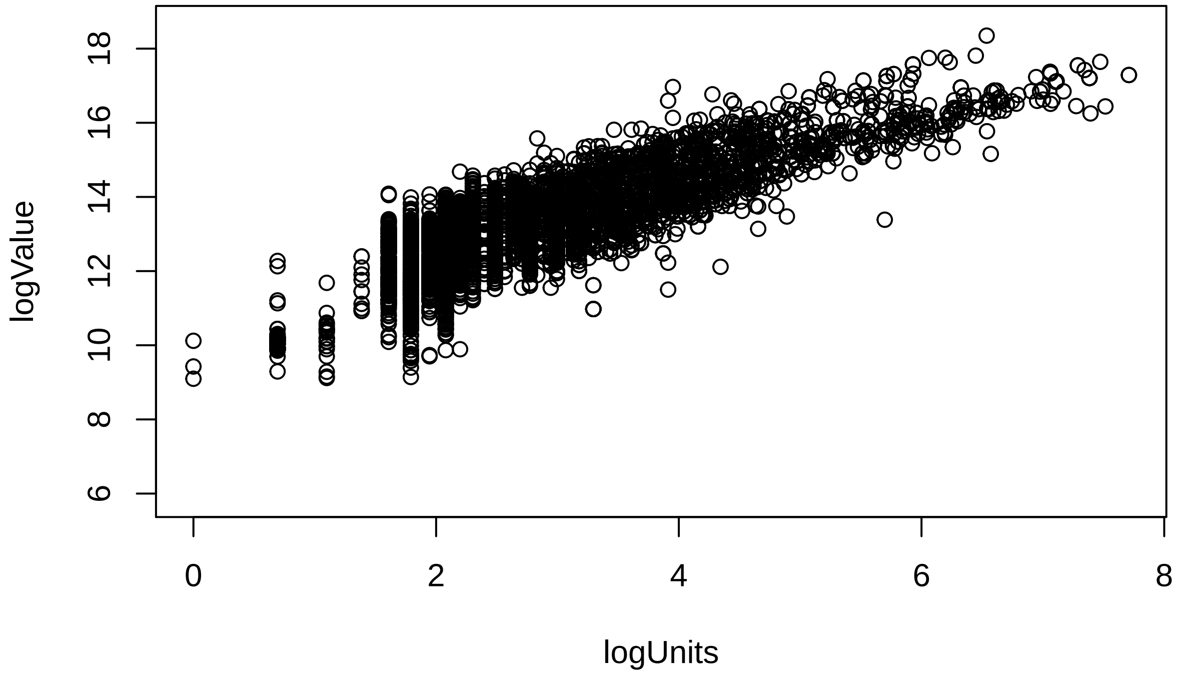
The correlation between property logValue and property logUnits in the whole data is 0.84. In just Manhattan it is 0.58. In just Brooklyn it is 0.67. For properties built after 2000 it is 0.478. For properties built before 2000 it is 0.477.

iv. Make two plots showing property logValue against property logUnits for Manhattan and Brooklyn.

```
#plots logValue vs logUnits for properties in Manhattan
plot(log(housing$res_units[manhat]),log(housing$assessed_value[manhat]),xlab="logUnits",ylab="logValue"
```



```
#plots logValue vs logUnits for properties in Brooklyn
plot(log(housing$res_units[brook]),log(housing$assessed_value[brook]),xlab="logUnits",ylab="logValue")
```



v. Consider the following block of code. Give a single line of R code which gives the same final answer as the block of code.

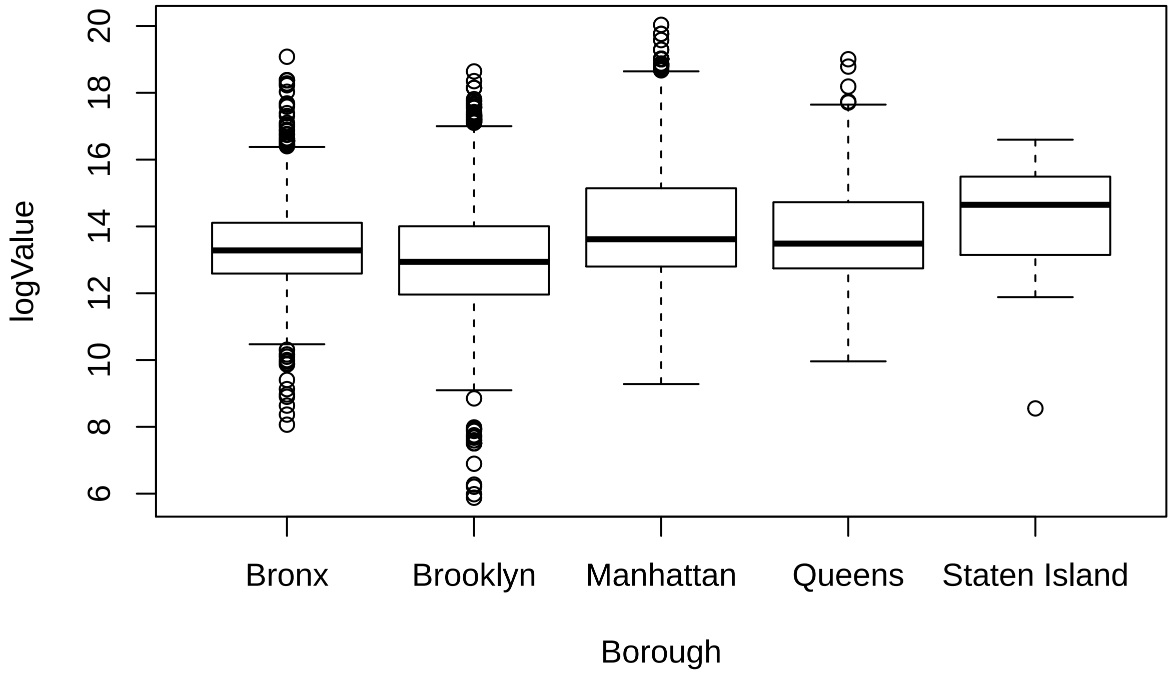
```r
manhat.props <- (1:dim(housing)[1])[housing$boro_name=="Manhattan"]
#stores index number of properties in Manhattan

med.value <- median(housing$assessed_value[manhat.props], na.rm=TRUE)
#calculates median value of properties in Manhattan
```

vi. Make side-by-side box plots comparing property logValue across the five boroughs

```r
#factorizes different boroughs from dataset
housing$boro_name <- factor(housing$boro_name)

#side-by-side box plots of logValue of each boroughs
boxplot(logValue ~ boro_name, data = housing, ylab = "logValue", xlab = "Borough")
```



vii. For five boroughs, what are the median property values?

```
bronx <- housing$boro_name=="Bronx" #filters so to keep only properties in Bronx
queens <- housing$boro_name=="Queens" #filters so to keep only properties in Queens
staten <- housing$boro_name=="Staten Island" #filter so to keep only properties in Staten Island

#calculates median property value  in the five boroughs.
median(housing$assessed_value[manhat], na.rm=TRUE)
```

```
## [1] 820350
```

```
median(housing$assessed_value[brook], na.rm=TRUE)
```

```
## [1] 416014
```

```
median(housing$assessed_value[bronx], na.rm=TRUE)
```

```
## [1] 587250
```

```
median(housing$assessed_value[queens], na.rm=TRUE)
```

```
## [1] 719100
```

```
median(housing$assessed_value[staten], na.rm=TRUE)
```

```
## [1] 2296350
```

The median property value in Manhattan in $820,350, in Brooklyn it is $416,014, in Bronx it is $587,250, in Queens it is $719,100 and in Staten Island it is $2,296,350.