Assignment 2

Tushar Ponkshe

September 27, 2018

Part 1

```
#i
housing <- read.csv("C:/Users/tusha/OneDrive/Desktop/Data/NYChousing.csv", as.is = TRUE, header
 = TRUE)

#ii
rows<-dim(housing)[1]; rows
```

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

```
cols<-dim(housing)[2]; cols # 2506 rows, 22 columns
```

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

```
#iii
col_sum <- apply(is.na(housing), 2, sum)
#The apply function gives the sum total or the total number of NA values accross each column of
 the dataset.

#iv
new_housing <- na.omit(housing)

#v
a<-dim(new_housing)[1]
b<-dim(housing)[1]
diff<-b-a
diff
```

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

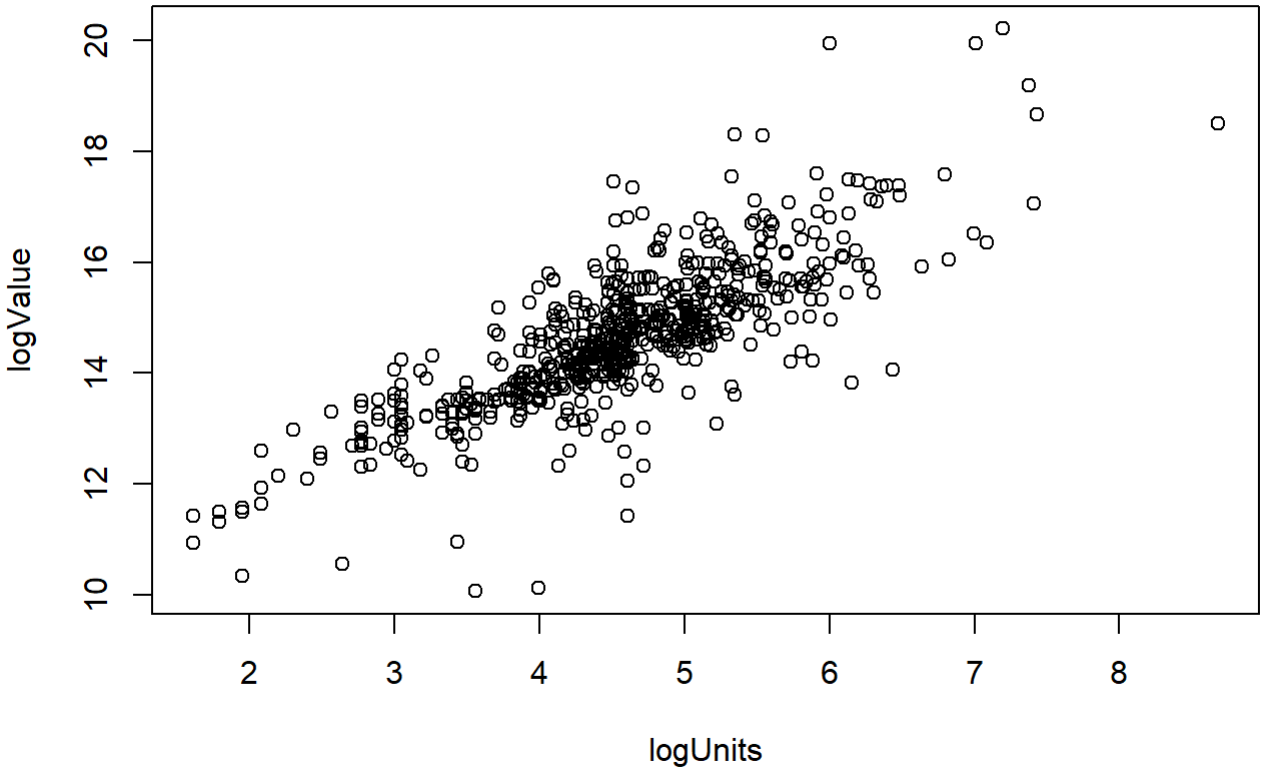
```
# Removed 1876 rows. This result is consistent with the result in iii

#vi
new_housing$logValue <- log(new_housing$Value)
summary(new_housing$logValue)
```

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

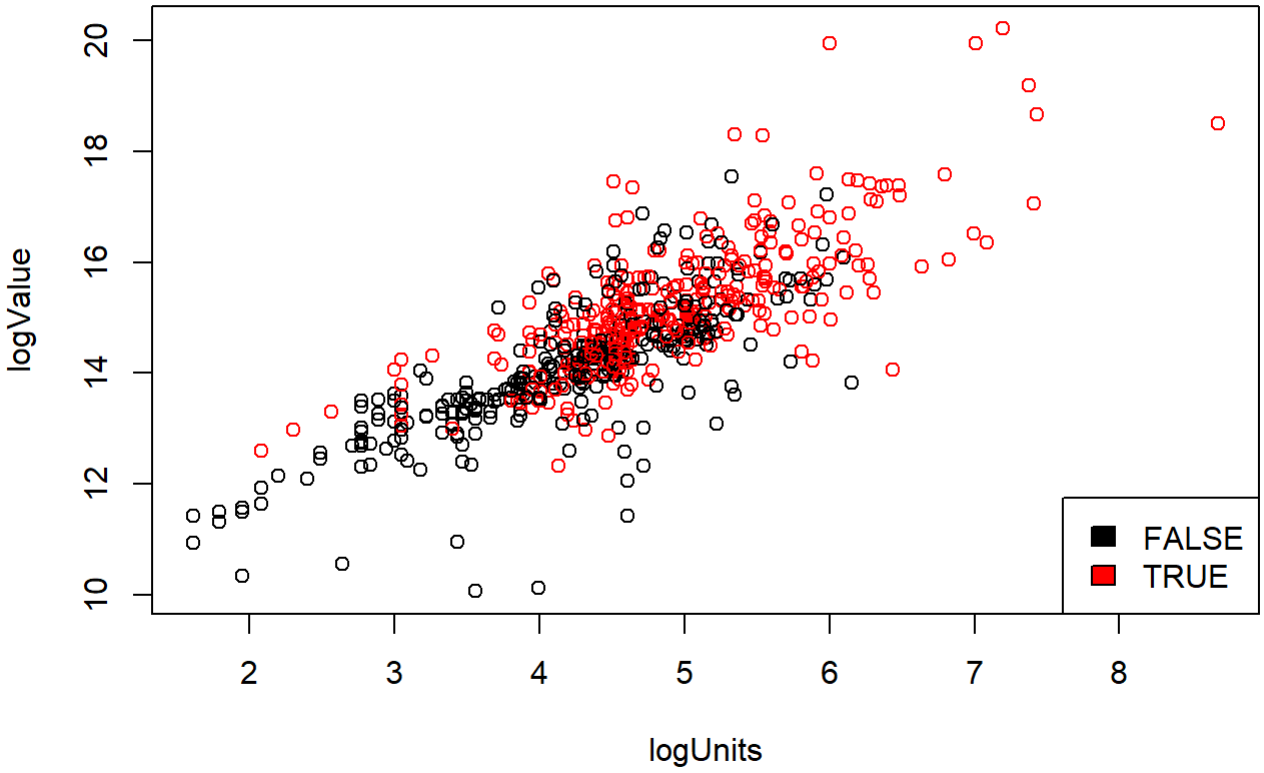
```
#Min = 10.06, mean = 14.65, max = 20.22

#vii
new_housing$logUnits <- log(new_housing$UnitCount)

#viii
new_housing$after1950 <- new_housing$YearBuilt >= 1950
```

Part 2

```
#i
plot(new_housing$logUnits, new_housing$logValue, xlab = 'logUnits', ylab = 'logValue')
```




```
#ii
plot(new_housing$logUnits, new_housing$logValue, xlab = 'logUnits', ylab = 'logValue', col = fac
tor(new_housing$after1950))
legend("bottomright", legend = levels(factor(new_housing$after1950)), fill = unique(factor(new_h
ousing$after1950)))
```




```
#The plot describes the log value of property based on the number of units in the property. We t
ake the log scale to account for the skewedness in the data.
cov(new_housing$logUnits, new_housing$logValue)
```

```
## [1] 0.9955796
```

```
#The covariance of the two variables in positive which shows the tendency in the linear relation
ship between the two variables
#The coloring in the plot tells us the relationship between logValue and logUnits before and aft
er the year 1950.

#iii
cor(new_housing$logValue, new_housing$logUnits)
```

```
## [1] 0.7988655
```

```
cor(new_housing$logValue[which(new_housing$Borough=='Manhattan')], new_housing$logUnits[which(ne
w_housing$Borough=='Manhattan')])
```

```
## [1] 0.8710823
```

```
cor(new_housing$logValue[which(new_housing$Borough=='Brooklyn')], new_housing$logUnits[which(new_housing$Borough=='Brooklyn')])
```

```
## [1] 0.8053241
```

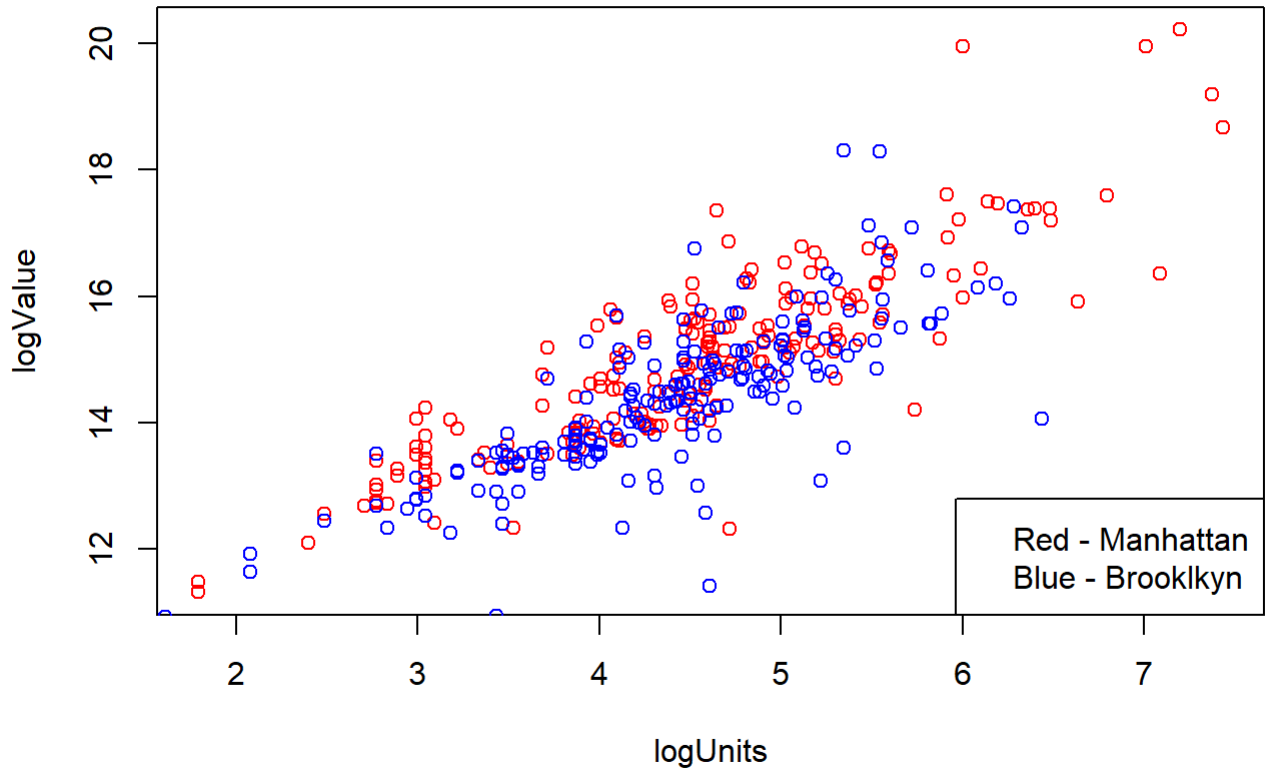
```
cor(new_housing$logValue[which(new_housing$after1950==TRUE)], new_housing$logUnits[which(new_housing$after1950==TRUE)])
```

```
## [1] 0.746731
```

```
cor(new_housing$logValue[which(new_housing$after1950==FALSE)], new_housing$logUnits[which(new_housing$after1950==FALSE)])
```

```
## [1] 0.7720285
```

```
#iv
plot(new_housing$logUnits[which(new_housing$Borough=='Manhattan')], new_housing$logValue[which(new_housing$Borough=='Manhattan')], col='red', xlab='logUnits', ylab = 'logValue')
points(new_housing$logUnits[which(new_housing$Borough=='Brooklyn')], new_housing$logValue[which(new_housing$Borough=='Brooklyn')], col='blue')
legend("bottomright", legend = c("Red - Manhattan","Blue - Brooklkyn"))
```



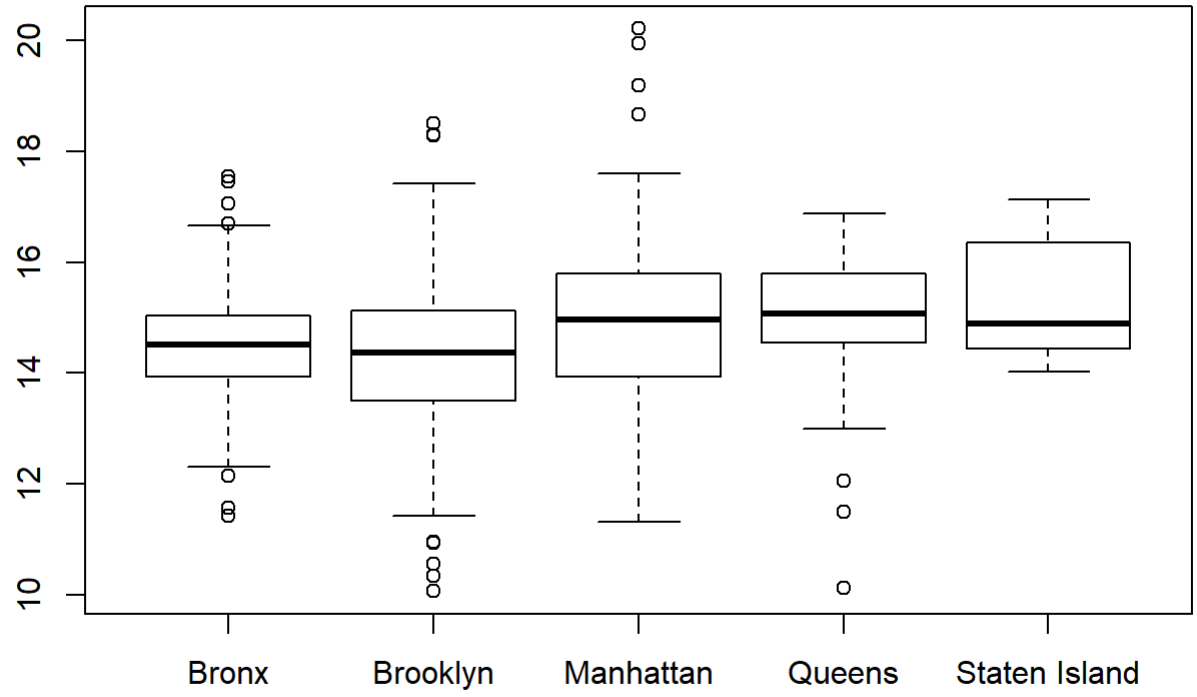
```
#v
med.value <- median(new_housing$Value[which(new_housing$Borough == 'Manhattan')])
med.value
```

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

```
#Compared to the output of the given code ->
manhat.props <- c()
for (props in 1:nrow(new_housing)) {
  if (new_housing$Borough[props] == "Manhattan") {
    manhat.props <- c(manhat.props, props)
  }
}
med.value <- c()
for (props in manhat.props) {
  med.value <- c(med.value, new_housing$Value[props])
}
med.value <- median(med.value, na.rm = TRUE)
med.value
```

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

```
#vi
boxplot(new_housing$logValue~new_housing$Borough)
```




```
#vii
median(new_housing$Value[which(new_housing$Borough=='Bronx')])
```

```
## [1] 2008260
```

```
median(new_housing$Value[which(new_housing$Borough=='Brooklyn')])
```

```
## [1] 1749465
```

```
median(new_housing$Value[which(new_housing$Borough=='Manhattan')])
```

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

```
median(new_housing$Value[which(new_housing$Borough=='Queens')])
```

```
## [1] 3529800
```

```
median(new_housing$Value[which(new_housing$Borough=='Staten Island')])
```

```
## [1] 2952900
```