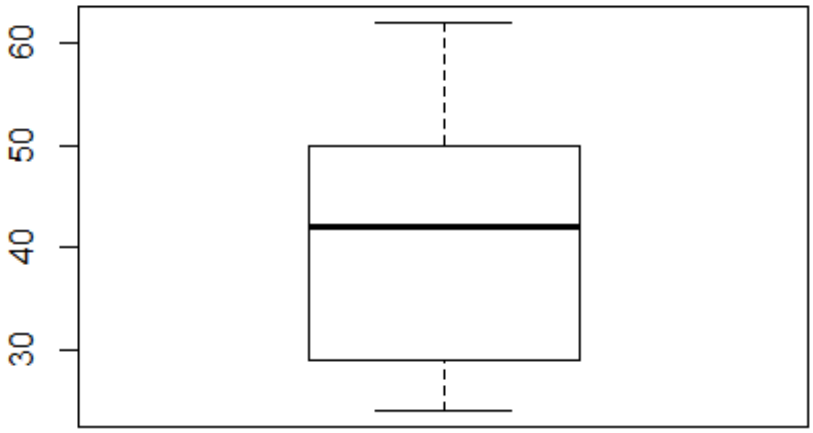
EDA_R_code.R

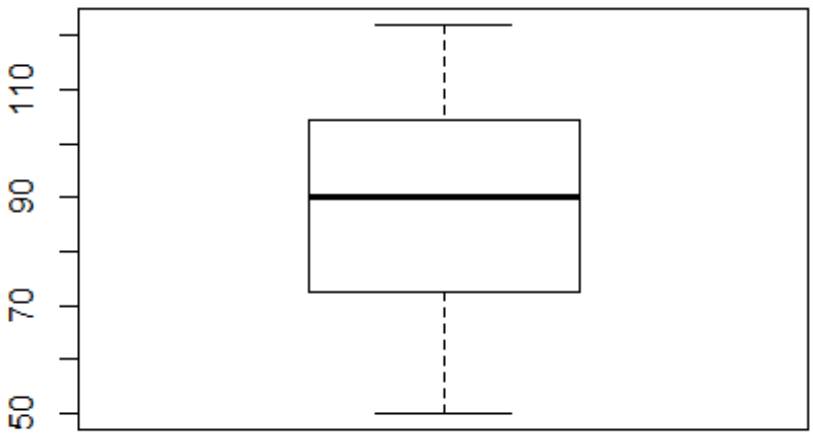
User02

Tue Mar 13 10:06:11 2018

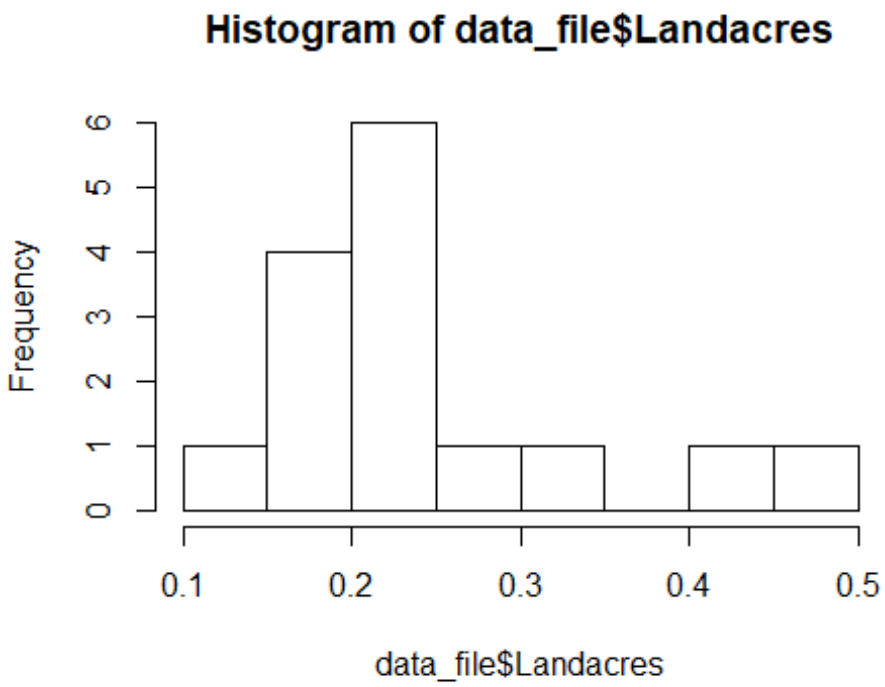
```
### Final ###
setwd("C:/Users/User02/Google Drive/Business Analytics/Business Analytics
Video/Moodle Upload/Explorartory Data Analytics/")
data_file <- read.csv("EDA_data.csv")
View(data_file)
summary(data_file)

##       Name        Age        Gender    Education     Salary
##  Bruce  :1   Min.   :24.00   F:7    Grad    :9   Min.   : 50.00
##  Carol  :1   1st Qu.:29.00   M:8    PostGrad:6   1st Qu.: 72.50
##  Donald :1   Median :42.00                       Median : 90.00
##  Harret :1   Mean   :40.33                       Mean   : 88.53
##  Henry  :1   3rd Qu.:50.00                       3rd Qu.:104.50
##  Jane   :1   Max.   :62.00                       Max.   :122.00
##  (Other):9
##  AppraisedValue         Location   Landacres      HouseSizesqrft
##  Min.   : 299.0   Glen Cove :5   Min.   :0.1377   Min.   :1120
##  1st Qu.: 390.4   Long Beach:5   1st Qu.:0.1732   1st Qu.:1707
##  Median : 517.7   Roslyn    :5   Median :0.2290   Median :2042
##  Mean   : 547.2                  Mean   :0.2425   Mean   :2141
##  3rd Qu.: 600.0                  3rd Qu.:0.2523   3rd Qu.:2472
##  Max.   :1200.0                  Max.   :0.4608   Max.   :4067
##
##      Rooms           Baths          Garage
##  Min.   :5.000   Min.   :1.000   Min.   :0.0
##  1st Qu.:6.750   1st Qu.:2.000   1st Qu.:0.0
##  Median :7.000   Median :2.000   Median :1.0
##  Mean   :7.167   Mean   :2.333   Mean   :0.8
##  3rd Qu.:8.000   3rd Qu.:2.750   3rd Qu.:1.0
##  Max.   :9.000   Max.   :4.000   Max.   :2.0
##  NA's   :3

boxplot(data_file$Age)
```



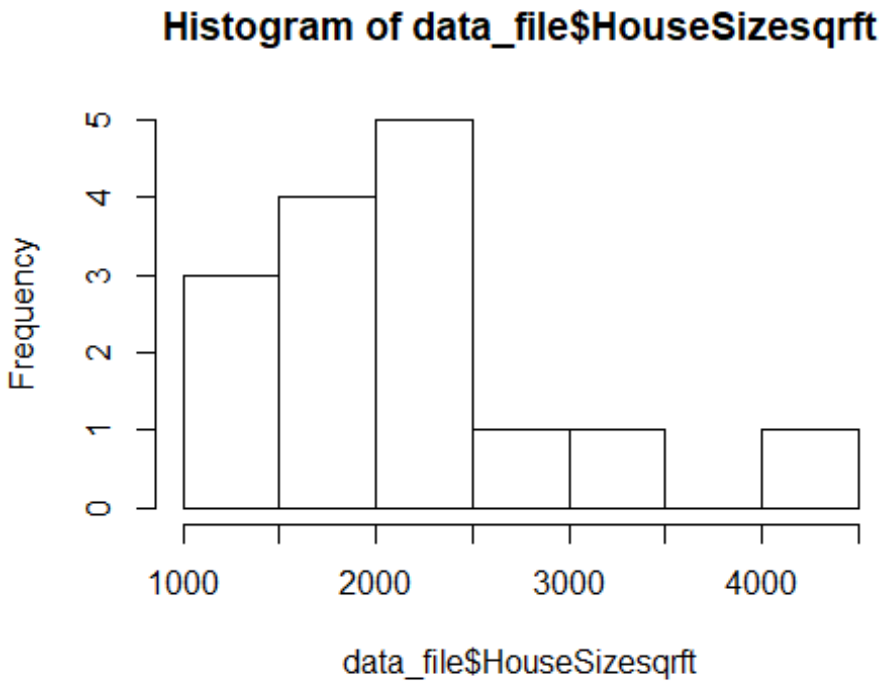

```r
boxplot(data_file$Salary)
```




```r
hist(data_file$Landacres)
```

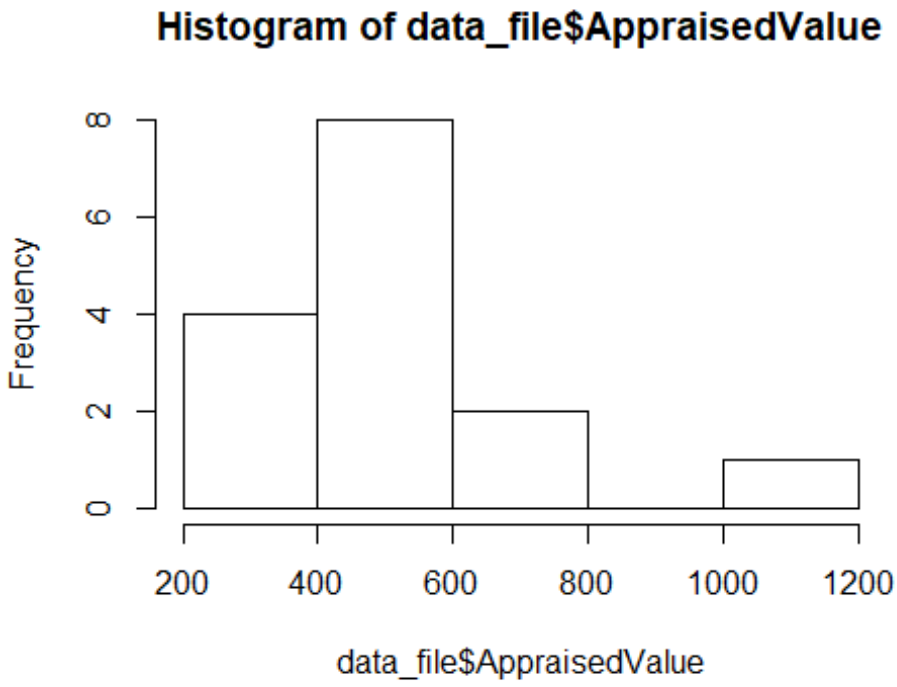



```r
hist(data_file$HouseSizesqrft)
```

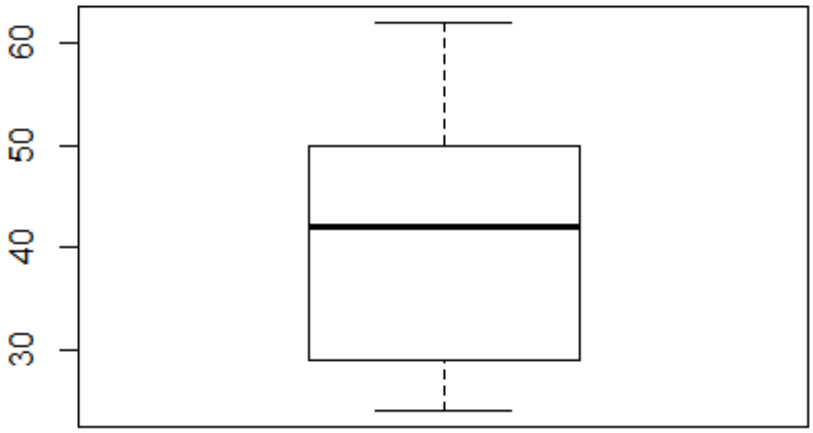



```r
hist(data_file$Salary)
```


```r
hist(data_file$AppraisedValue)
```

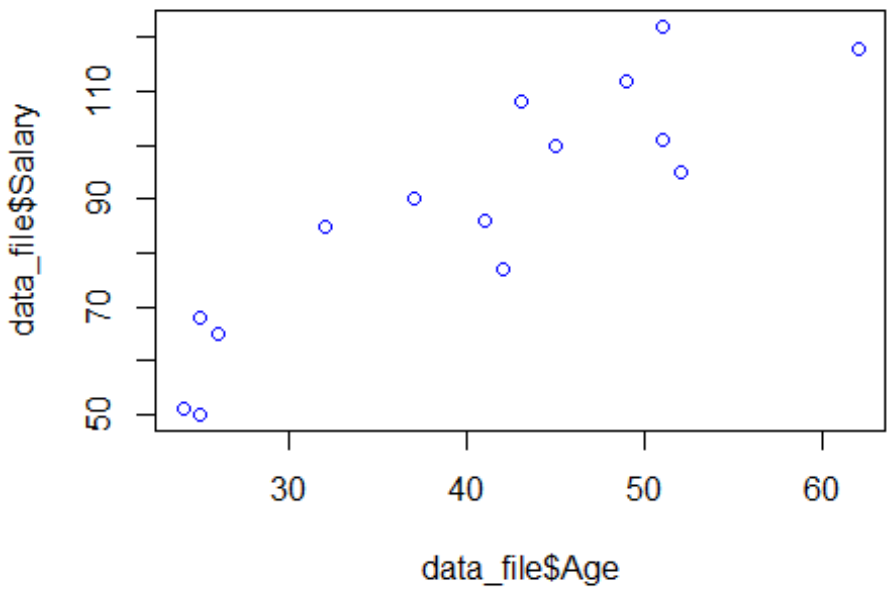



```r
boxplot(data_file$Age)
```




```r
hist(data_file$Salary)
```


```r
plot(data_file$Age, data_file$Salary, col = "blue")
```




```
#Two-way table
counts = table(data_file$Education,data_file$Gender)
counts
```

```
##
##            F M
##   Grad     3 6
##   PostGrad 4 2
```

```
#Stacked column chart
barplot(counts, main = "Data distribution by Education Vs Gender",col =
c("blue","red"),legend=rownames(counts),
        args.legend = list(x = "bottom", bty = "n", inset=c(-0.40, -.40)))
```

```
#Imputing with Mean
data_file$Rooms[is.na(data_file$Rooms)] <- mean(data_file$Rooms, na.rm =
TRUE)
View(data_file)
summary(data_file)
```

```
##       Name        Age        Gender    Education     Salary
##  Bruce  :1   Min.   :24.00   F:7    Grad    :9   Min.   : 50.00
##  Carol  :1   1st Qu.:29.00   M:8    PostGrad:6   1st Qu.: 72.50
##  Donald :1   Median :42.00                       Median : 90.00
##  Harret :1   Mean   :40.33                       Mean   : 88.53
##  Henry  :1   3rd Qu.:50.00                       3rd Qu.:104.50
```

```
##  Jane   :1   Max.   :62.00                       Max.   :122.00
##  (Other):9
##  AppraisedValue         Location   Landacres      HouseSizesqrft
##  Min.   : 299.0   Glen Cove :5   Min.   :0.1377   Min.   :1120
##  1st Qu.: 390.4   Long Beach:5   1st Qu.:0.1732   1st Qu.:1707
##  Median : 517.7   Roslyn    :5   Median :0.2290   Median :2042
##  Mean   : 547.2                  Mean   :0.2425   Mean   :2141
##  3rd Qu.: 600.0                  3rd Qu.:0.2523   3rd Qu.:2472
##  Max.   :1200.0                  Max.   :0.4608   Max.   :4067
##
##      Rooms           Baths           Garage
##  Min.   :5.000   Min.   :1.000   Min.   :0.0
##  1st Qu.:7.000   1st Qu.:2.000   1st Qu.:0.0
##  Median :7.167   Median :2.000   Median :1.0
##  Mean   :7.167   Mean   :2.333   Mean   :0.8
##  3rd Qu.:8.000   3rd Qu.:2.750   3rd Qu.:1.0
##  Max.   :9.000   Max.   :4.000   Max.   :2.0
##
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

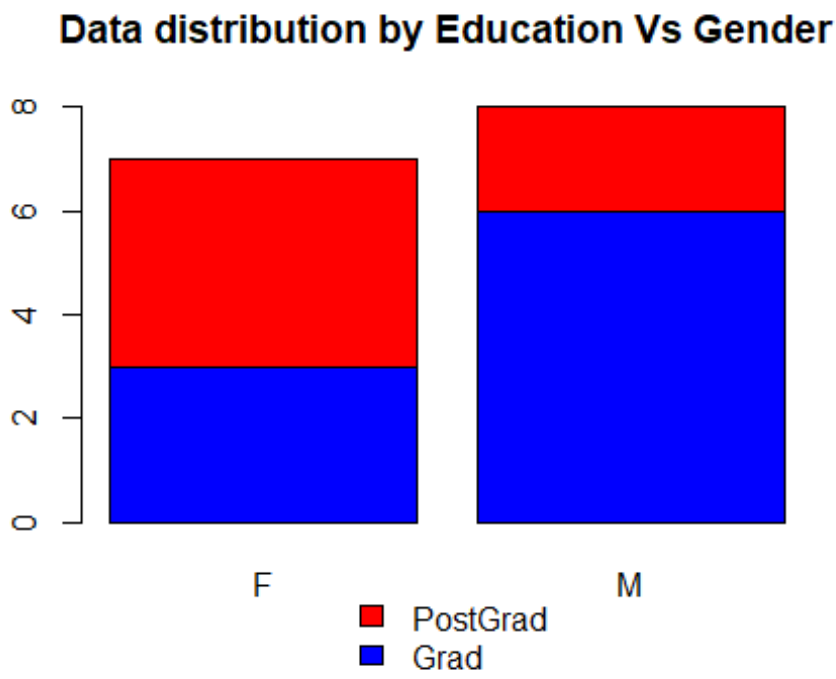
```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```



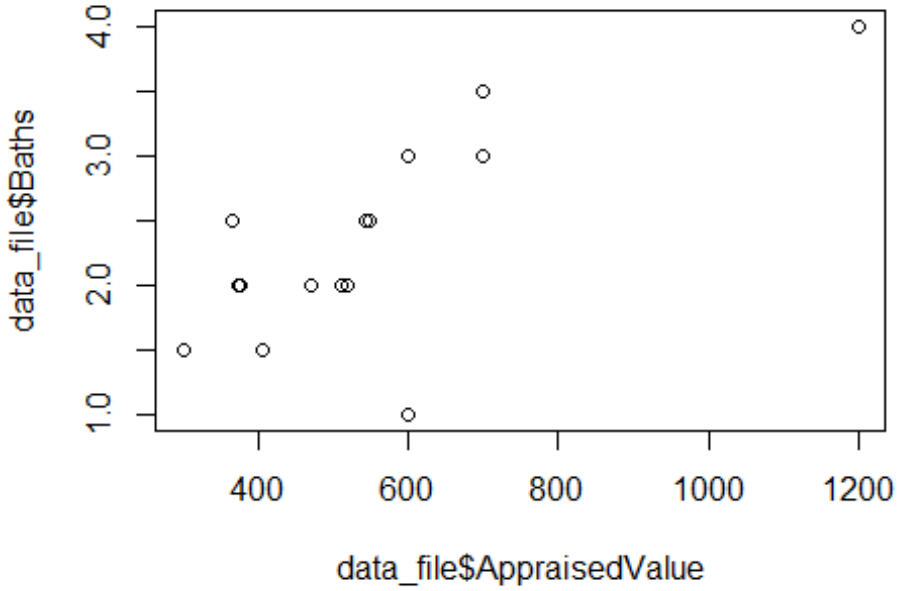

```
impute(data_file$Rooms, mean) # replace with mean
```

```
##  [1] 8.000000 7.000000 7.000000 8.000000 7.166667 7.000000 8.000000
##  [8] 5.000000 6.000000 7.166667 7.000000 9.000000 8.000000 7.166667
## [15] 6.000000
```

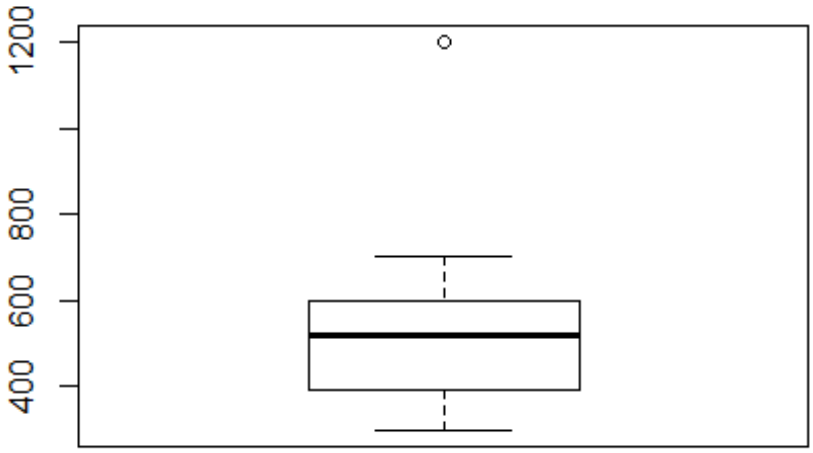
```
impute(data_file$Rooms, median) #  replace with median
```

```
##  [1] 8.000000 7.000000 7.000000 8.000000 7.166667 7.000000 8.000000
##  [8] 5.000000 6.000000 7.166667 7.000000 9.000000 8.000000 7.166667
## [15] 6.000000
```

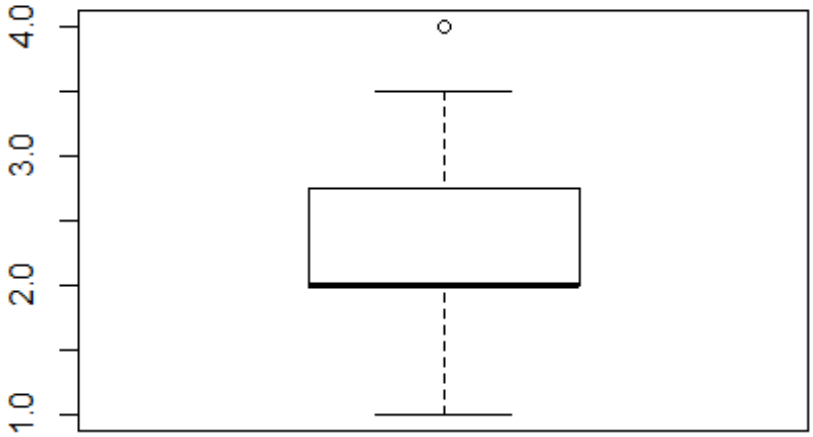
```
plot(data_file$AppraisedValue,data_file$Baths)
```



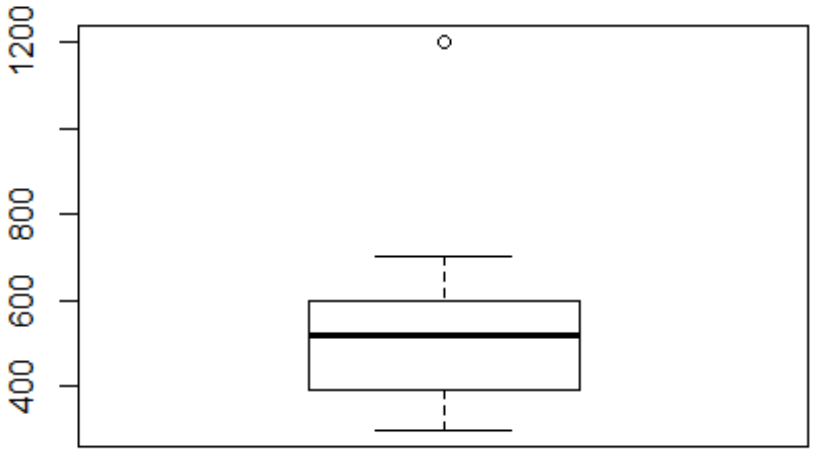

```
boxplot(data_file$AppraisedValue)
```




```
boxplot(data_file$Baths)
```




```
boxplot(data_file$AppraisedValue)
```




```
summary(data_file$AppraisedValue)
```

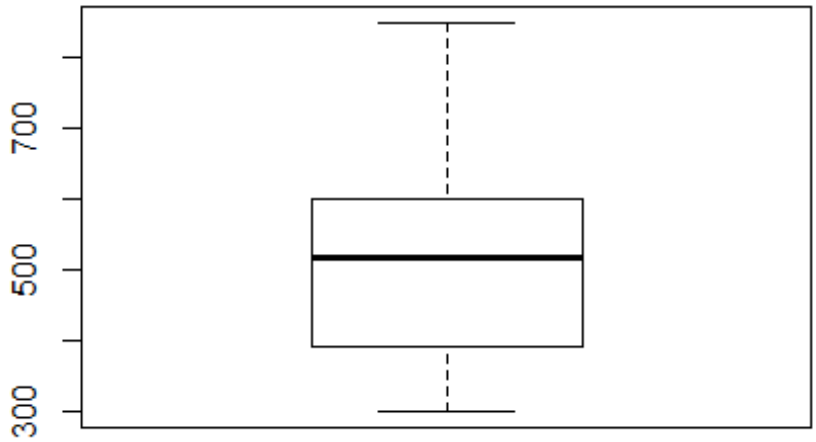
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   299.0   390.4   517.7   547.2   600.0  1200.0
```

```
quantile(data_file$AppraisedValue, .95)
```

```
## 95%
## 850
```

```
data_file$AppraisedValue = ifelse(data_file$AppraisedValue >= 1000, 850,
data_file$AppraisedValue)
```

```
boxplot(data_file$AppraisedValue)
```




```
summary(data_file$AppraisedValue)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   299.0   390.4   517.7   523.9   600.0   850.0
```

```
quantile(data_file$AppraisedValue, .95) #95th percentile
```

```
## 95%
## 745
```

```
quantile(data_file$AppraisedValue, .05) #5th percentile
```

```
##    5%
## 344.5
```

```r
#Dummy var creation
library(lme4)

## Loading required package: Matrix

new = dummy(data_file$Location)

new_data = cbind(data_file,new)
View(new_data)
```