In [1]:
```
data <- read.csv("mydata.csv", sep=";", dec=",")
```

First Task

In [33]:
```
summary(data)
```

```
  AddressCount    CallsCount     ClicksCount       FirmsCount
 Min.   :   9   Min.   :   20   Min.   :   258   Min.   :  14.0
 1st Qu.:  81   1st Qu.:  346   1st Qu.:  2055   1st Qu.:  71.5
 Median : 371   Median :  931   Median :  6921   Median : 185.0
 Mean   :1048   Mean   : 3649   Mean   : 21826   Mean   : 305.1
 3rd Qu.:1195   3rd Qu.: 2458   3rd Qu.: 30626   3rd Qu.: 402.5
 Max.   :9552   Max.   :48497   Max.   :167155   Max.   :2379.0
    GeoPart          MobilePart       UsersCount       Distance
 Min.   :0.09292   Min.   :0.0900   Min.   :  157   Min.   : 714.8
 1st Qu.:0.28153   1st Qu.:0.3573   1st Qu.: 1168   1st Qu.:1562.1
 Median :0.32234   Median :0.4637   Median : 2934   Median :2586.5
 Mean   :0.34264   Mean   :0.4457   Mean   : 9753   Mean   :2669.4
 3rd Qu.:0.41691   3rd Qu.:0.5517   3rd Qu.:13265   3rd Qu.:3575.7
 Max.   :0.55618   Max.   :0.7373   Max.   :61127   Max.   :6292.2
     IsGeo
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.3544
 3rd Qu.:1.0000
 Max.   :1.0000
```

In [37]:
```
apply(data, 2, var)
```

AddressCount: 2696381.13956508 **CallsCount:** 66001088.5780591 **ClicksCount:** 1054622995.39727 **FirmsCount:** 145963.799740344 **GeoPart:** 0.0107345357860208 **MobilePart:** 0.0213541594853761 **UsersCount:** 193969566.086336 **Distance:** 2038570.87377841 **IsGeo:** 0.2317429406037

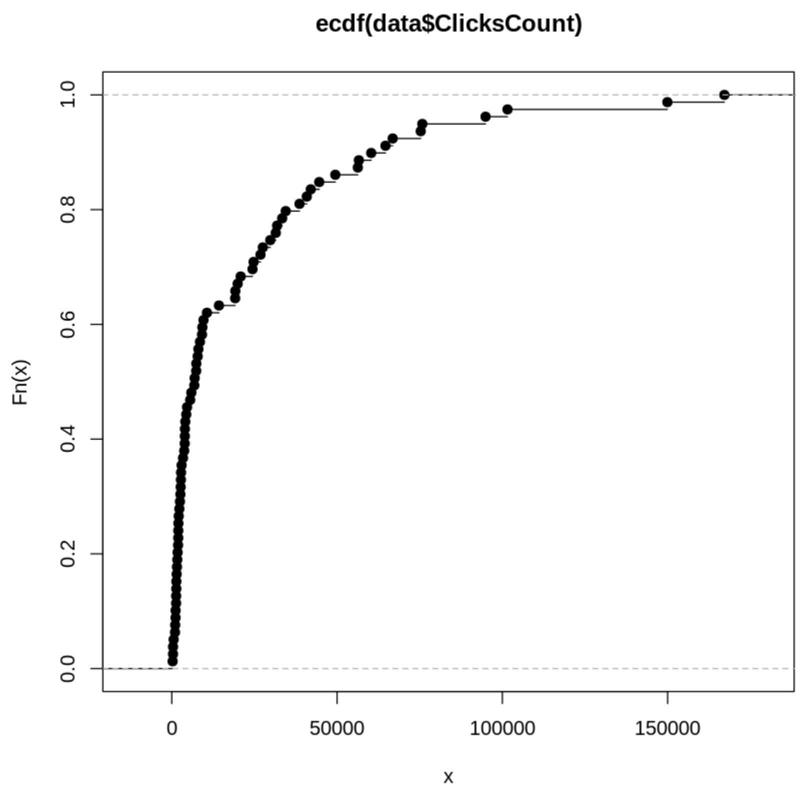
In [36]:
```
apply(data, 2, sd)
```

AddressCount: 1642.06611912099 **CallsCount:** 8124.10540170787 **ClicksCount:** 32474.9595134047 **FirmsCount:** 382.052090349397 **GeoPart:** 0.103607604865767 **MobilePart:** 0.146130624734777 **UsersCount:** 13927.2957205028 **Distance:** 1427.78530381091 **IsGeo:** 0.481396863932141

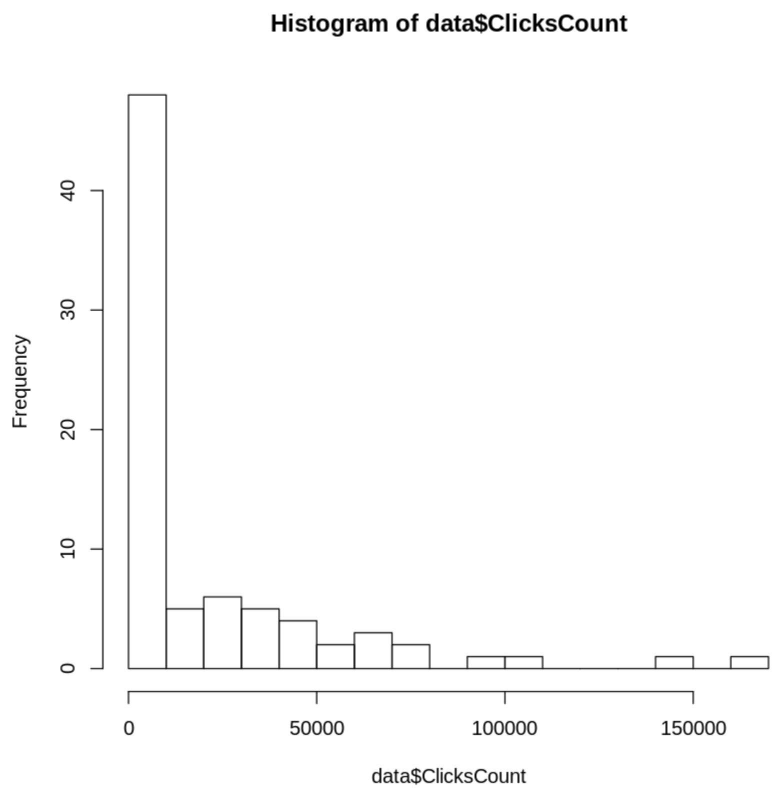
Second Task

Lets analyse Clicks Count

In [78]:
```
plot(ecdf(data$ClicksCount))
```

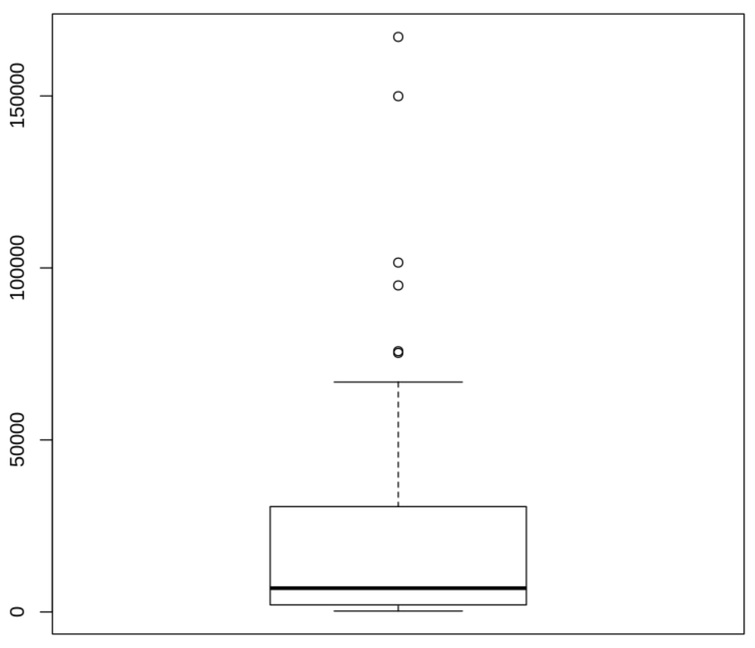


In [80]:
```
hist(data$ClicksCount, breaks="FD")
```



Based on the histogram, we see that our data has a LogN distribution. But we have quite a few outliers on the right, so it's too hard too bee sure.

In [85]:
```
boxplot(data$ClicksCount)
```



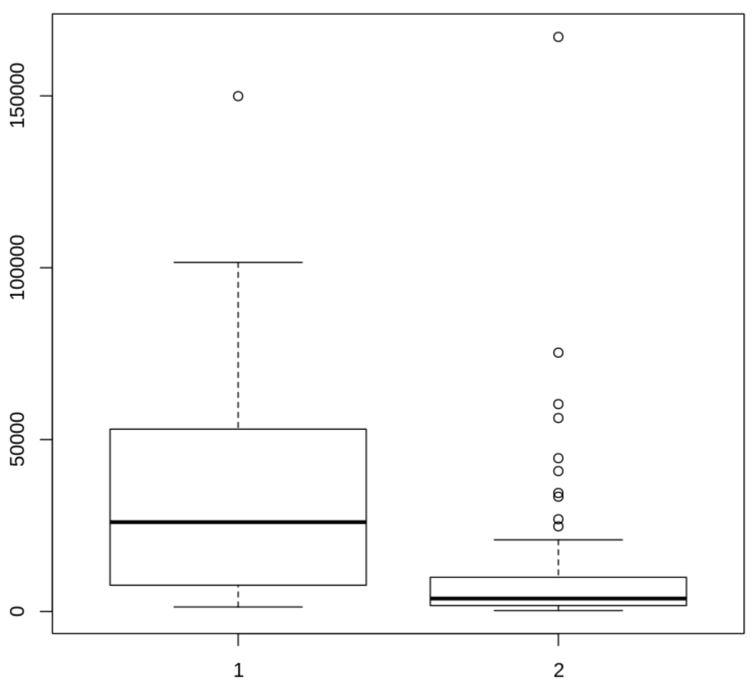
On boxplot we can see a lot of outliers. So, they can be part of a our distribution. Perhaps this is the reason that we mixed geo-dependent and geo-independent data

Third Task

In [88]:
```
geoData <- data[data$IsGeo == 1,]
```

In [89]:
```
notGeoData <- data[data$IsGeo == 0,]
```

In [90]:
```
boxplot(geoData$ClicksCount, notGeoData$ClicksCount)
```



Now we can see differance beetwen geo-dependent and geo-independent data. In geo-independent data, we see one outlier. Maybe we should exclude it from the data set. Actually we can see one huge outlier in geo-independent data, but other outliers of second set probably are part of our data set and we shoudn't exclude it