In [ ]:
```
data<-read.csv("_6c939ae474f29a209ab5e3901864d38a_Data_Projects.csv", sep=";", dec=".")
```

In [ ]:
```
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    
0,0929166666666667: 1   0,09              : 1   Min.   :  157  
0,137857900318134 : 1   0,133974358974359: 1   1st Qu.: 1168  
0,151738923296808 : 1   0,139612188365651: 1   Median : 2934  
0,187886279357231 : 1   0,175525339925834: 1   Mean   : 9753  
0,193484698914116 : 1   0,200644166213028: 1   3rd Qu.:13265  
0,203374777975133 : 1   0,204181869211339: 1   Max.   :61127  
(Other)           :73   (Other)           :73                  
            Distance       IsGeo       
1004,78676794652: 1   Min.   :0.0000  
1033,11276489631: 1   1st Qu.:0.0000  
1234,54344844949: 1   Median :0.0000  
1421,72399039962: 1   Mean   :0.3544  
1423,37651183958: 1   3rd Qu.:1.0000  
1437,3055534143 : 1   Max.   :1.0000  
(Other)         :73                   
```

Рассчитайте основные статистики (меры центра и меры разброса) по распределениям всех переменных, имеющихся в файле данных.

Там, где имеетмя (Other) меры не могли быть расчитаны.

In [ ]:
```
mean(data$AddressCount)
var(data$AddressCount)
```

1048.03797468354

2696381.13956508

In [ ]:
```
mean(data$CallsCount)
var(data$CallsCount)
```

3648.6835443038

66001088.5780591

In [ ]:
```
mean(data$ClicksCount)
var(data$ClicksCount)
```

21826.0126582278

1054622995.39727

In [ ]:
```
mean(data$FirmsCount)
var(data$FirmsCount)
```

305.088607594937

145963.799740344

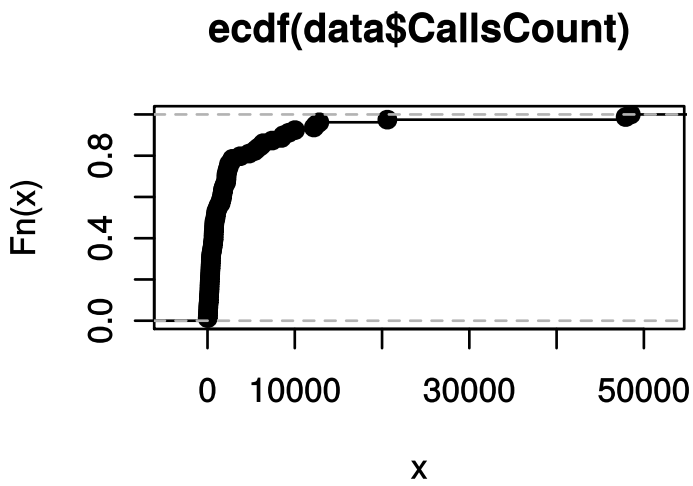
In [ ]:
```
mean(data$UsersCount)
var(data$UsersCount)
```

9753.12658227848

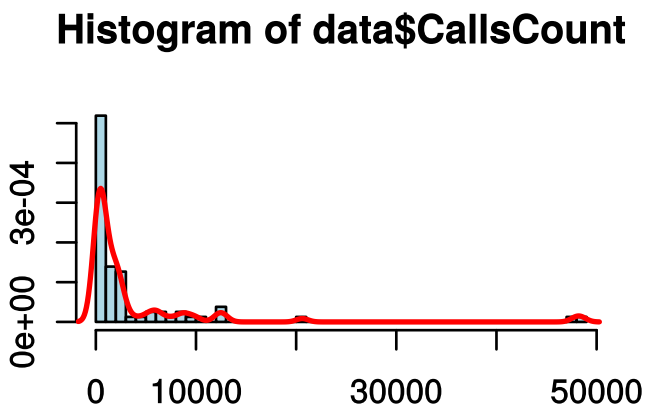
193969566.086336

2. Выберите наиболее интересный для вас количественный признак и охарактеризуйте его распределение при помощи соответствующих описательных статистик и графиков:

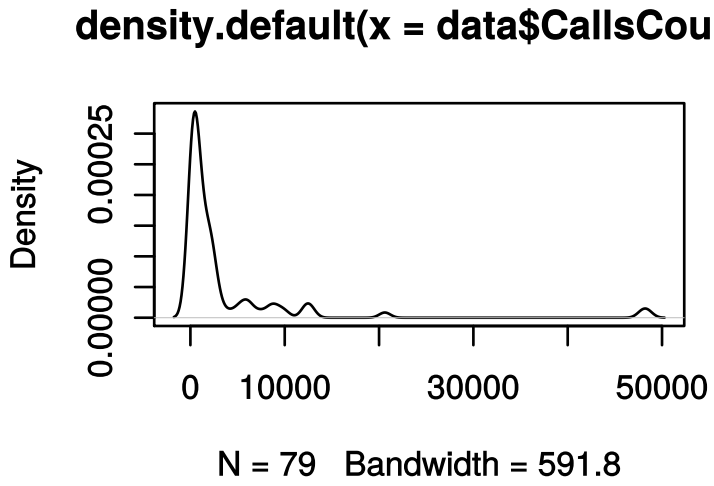
In [ ]:
```
plot(ecdf(data$CallsCount))
```



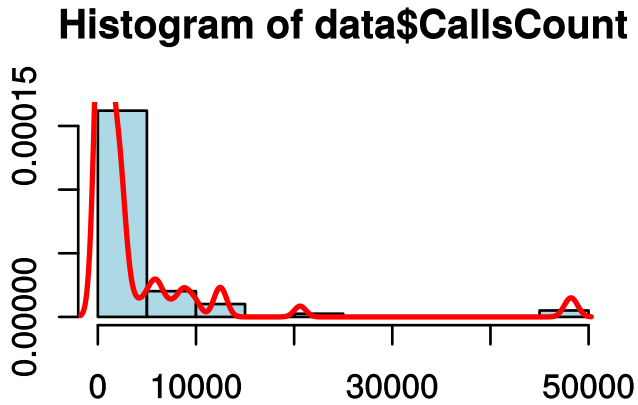
In [ ]:
```
hist(data$CallsCount, breaks = "FD", freq = FALSE, col = "lightblue", xlab = "Время", ylab = "Концент
lines(density(data$CallsCount), col = "red", lwd = 2)
```



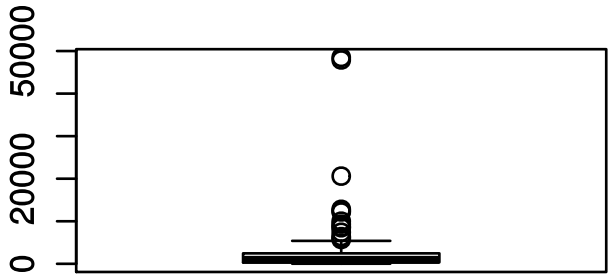
In [ ]:
```
plot(density(data$CallsCount))
```



In [ ]:
```
hist(data$CallsCount,freq = FALSE, col = "lightblue", xlab = "Время", ylab = "Концентрация")
lines(density(data$CallsCount), col = "red", lwd = 2)
```



In [ ]:
```
boxplot(data$CallsCount)
```



Вывод:

2.1. Какова форма распределения признака? - В выборке имется смесь распредлеений

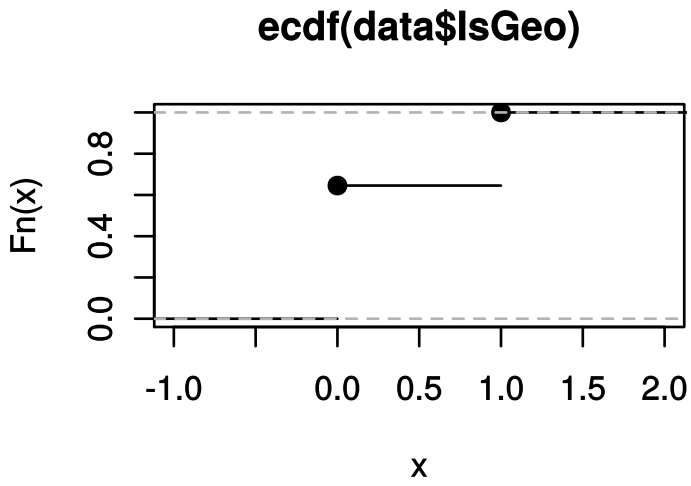
2.2. Можно ли говорить о том, что распределение признака согласуется с каким-либо теоретическим законом распределения? - Нет, на основании выбоки нельзя судить об едином распределении данных. Большинство данных сгруппированы до значения 15 тыс. Данные до 5 тыс. могут классифицироваться как данные с псевдонормальным распредленем и ассиметрией.

2.3. Есть ли в данных «аномалии» / «выбросы»? - Да, в данных есть выбросы (на основании BoxPlot)

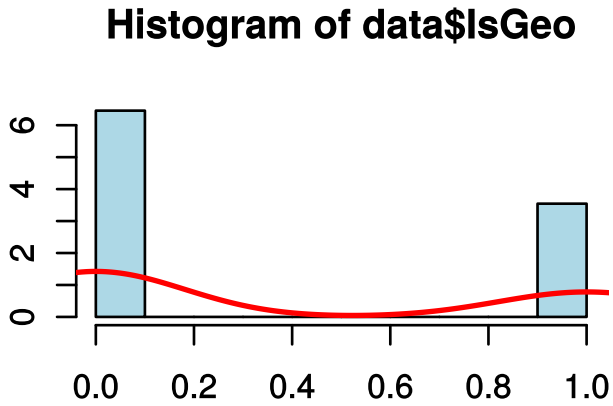
2.4. Какие меры центра и вариативности подходят для описания распределений лучше всего? Почему? - Ввиду того, что распеределение иммет выбросы и не близко к нормальному, то для опредления центра наиболее приемлимыми будут медиана и мода; для оценки вариативности - размах, межквартильный размахю

3. Сравните геозависимые и геонезависимые сферы (экспертная разметка, переменная IsGeo) по выбранному для анализа признаку. Есть ли отличия? В чем они состоят? (для ответа на вопрос используйте статистические и графические инструменты).

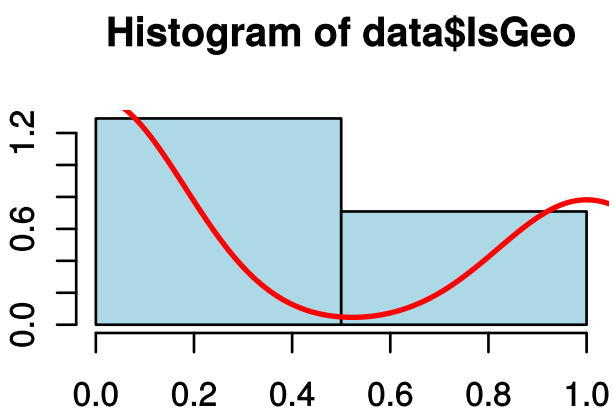
In [ ]:
```
plot(ecdf(data$IsGeo))
```



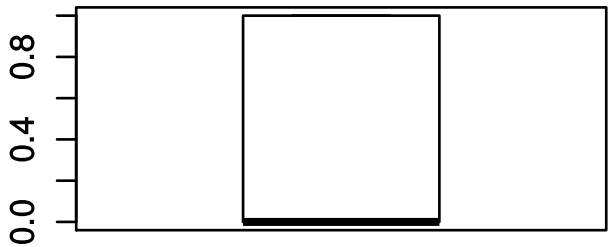
In [ ]:
```
hist(data$IsGeo,  freq = FALSE, col = "lightblue", xlab = "Время", ylab = "Концентрация")
lines(density(data$IsGeo), col = "red", lwd = 2)
```



In [ ]:
```
hist(data$IsGeo, breaks = "FD", freq = FALSE, col = "lightblue", xlab = "Время", ylab = "Концентраци
lines(density(data$IsGeo), col = "red", lwd = 2)
```



In [ ]:
```
boxplot(data$IsGeo)
```



Вывод: переменная IsGeo является бинарной номинальной переменной. Мерой среднего в таком случае будет выступать мода, мерой вариативности - отсутсвует.