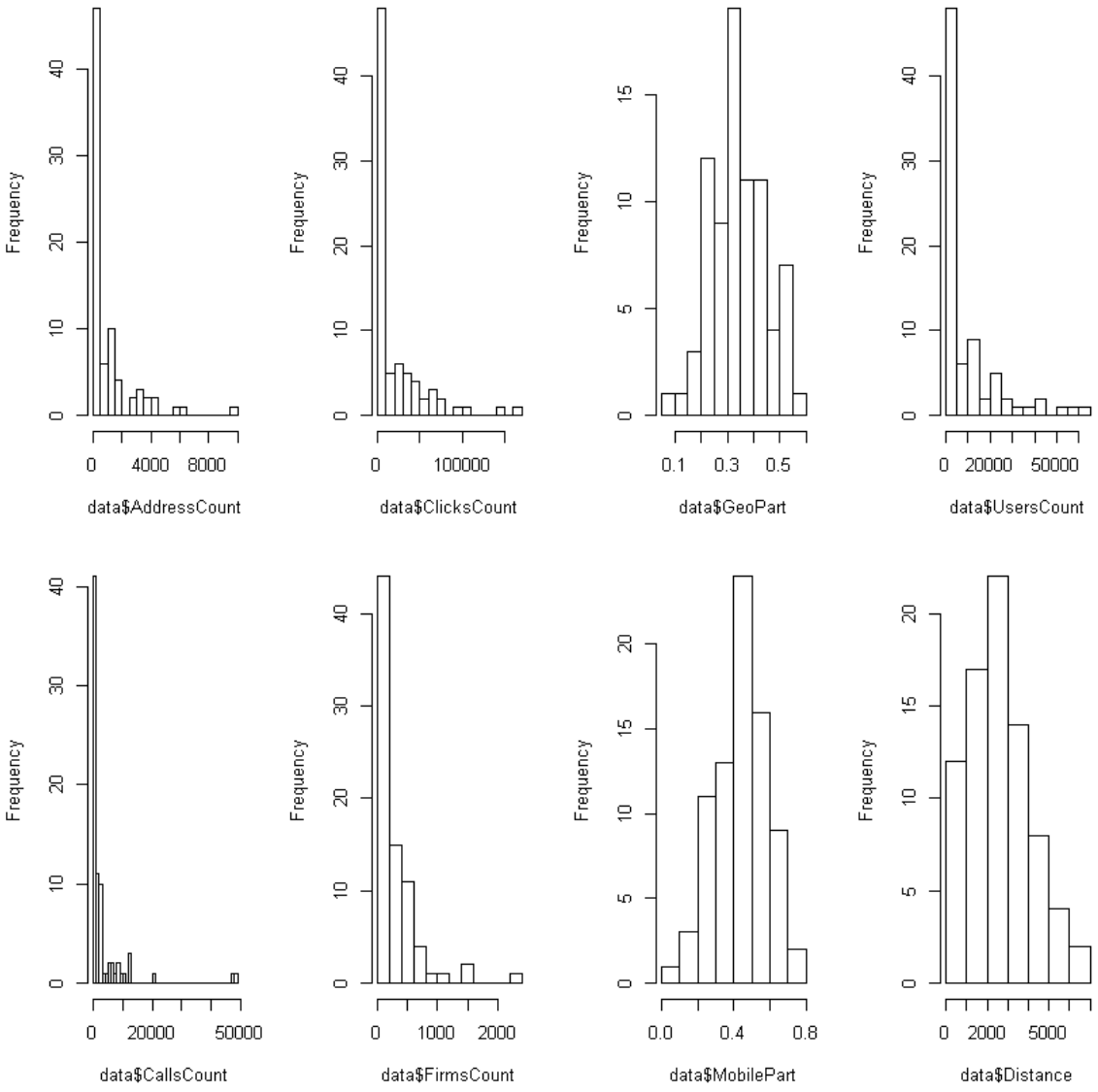
In [1]:
```
data<-read.csv("Data_Projects.csv", sep=";", dec=",")
```

1. Проанализируйте распределения признаков, которые вы хотите включить в модель. Обращайте внимание на наличие выбросов, не забывайте, что нетипичные значения всегда образуют отдельные, не информативные и ненаполненные кластеры.

In [2]:
```
head(data)
```

AddressCount	CallsCount	ClicksCount	FirmsCount	GeoPart	MobilePart	UsersCount	Distance
156	20	1903	176	0.4161044	0.5357625	1125	749.9661
17	37	258	20	0.2116788	0.4306569	157	2289.0324
78	56	1956	185	0.3494754	0.4765940	1195	1423.3765
14	70	378	19	0.3187184	0.4637437	206	3396.5661
111	90	4089	90	0.5561755	0.4905733	2934	1576.5142
53	96	1669	162	0.3989782	0.4217371	991	2337.6038

In [3]:
```
par(mfcol=c(2,4), mar=c(4,4,2,2))
hist(data$AddressCount, breaks="FD", main="")
hist(data$CallsCount, breaks="FD", main="")
hist(data$ClicksCount, breaks="FD", main="")
hist(data$FirmsCount, breaks="FD", main="")
hist(data$GeoPart, breaks="FD", main="")
hist(data$MobilePart, breaks="FD", main="")
hist(data$UsersCount, breaks="FD", main="")
hist(data$Distance, breaks="FD", main="")
```



Можно невооруженным глазом подметить, что AddressCount, CallsCount, ClicksCount, FirmsCount, UsersCount близки к логнормальному распределению. Тогда как GeoPart, MobilePart, Distance имеют куполообразное распределение с некоторой асимметрией и признаками смеси в данных. Distance явно асимметричен для нормального. Заранее известно, что в данных может быть смесь геонезависимых и геозависимых сфер, определенных бинарным признаком IsGeo (его гистограмму строить не имеет смысла). Все признаки (за исключением IsGeo) представлены в метрических шкалах, поэтому они все пригодны для дальнейшего анализа стандартными методами. Однако логнормально-распределенные признаки следует приблизить к нормальности, чтобы избежать смещенных оценок, опирающихся на меры среднего. Для этого прологарифмируем AddressCount, CallsCount, ClicksCount, FirmsCount, UsersCount для дальнейшего анализа.

2. Подготовьте переменные для включения в модель, зафиксируйте (и опишите в работе), как преобразовывались признаки, какие значения отбрасывались.

Прологарифмируем AddressCount, CallsCount, ClicksCount, FirmsCount, UsersCount для дальнейшего анализа.

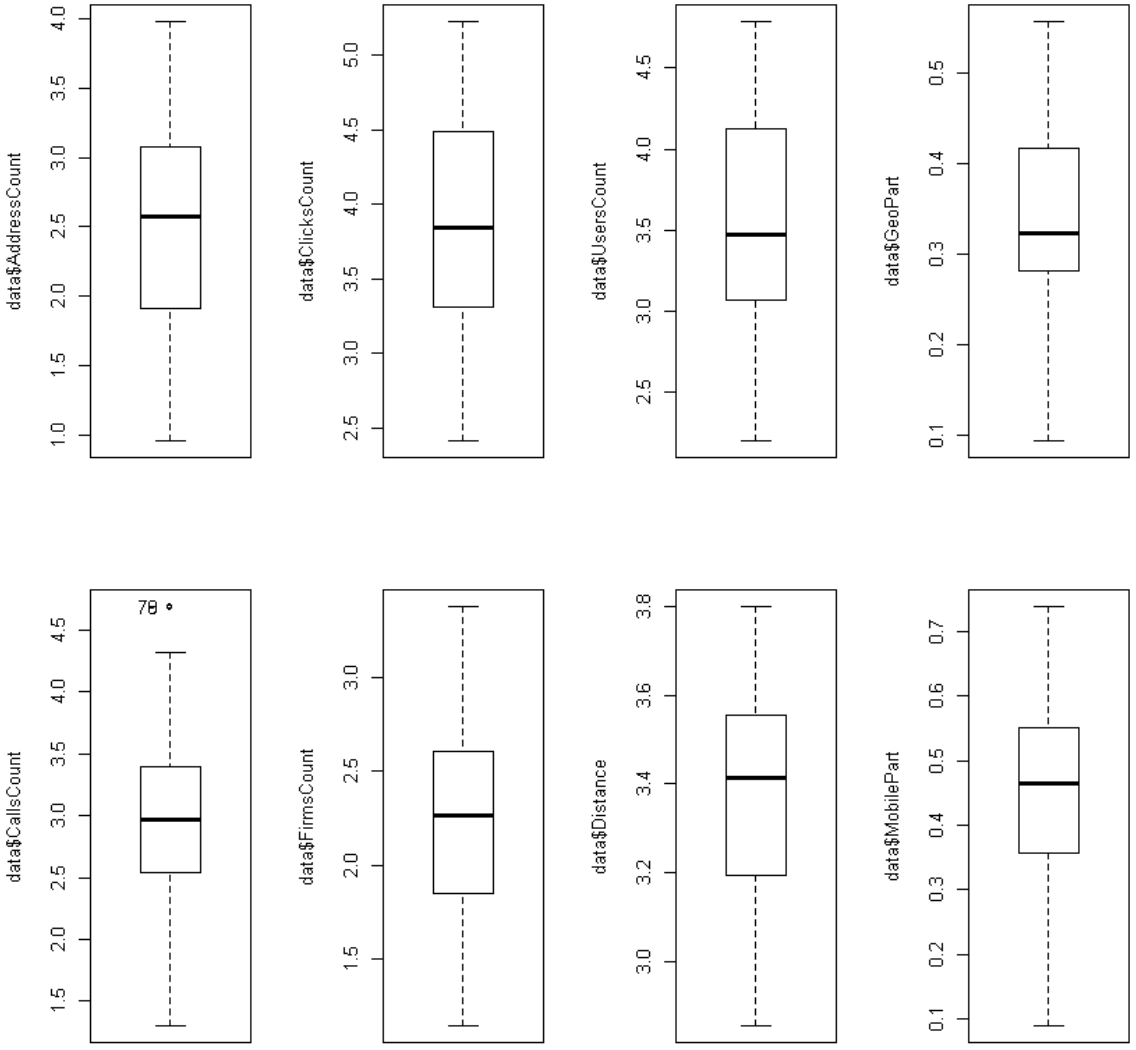
In [4]:
```
data$AddressCount<-log10(data$AddressCount)
data$CallsCount<-log10(data$CallsCount)
data$ClicksCount<-log10(data$ClicksCount)
data$FirmsCount<-log10(data$FirmsCount)
data$UsersCount<-log10(data$UsersCount)
data$Distance<-log10(data$Distance)
```

In [5]:
```
library(car)
```

```
Warning message:
"package 'car' was built under R version 3.6.3"Loading required package: carDat
a
```

In [6]:
```
par(mfcol=c(2,4), mar=c(4,4,2,2))
Boxplot(data$AddressCount)
Boxplot(data$CallsCount)
Boxplot(data$ClicksCount)
Boxplot(data$FirmsCount)
Boxplot(data$UsersCount, id=TRUE)
Boxplot(data$Distance)
Boxplot(data$GeoPart)
Boxplot(data$MobilePart)
```

78 79



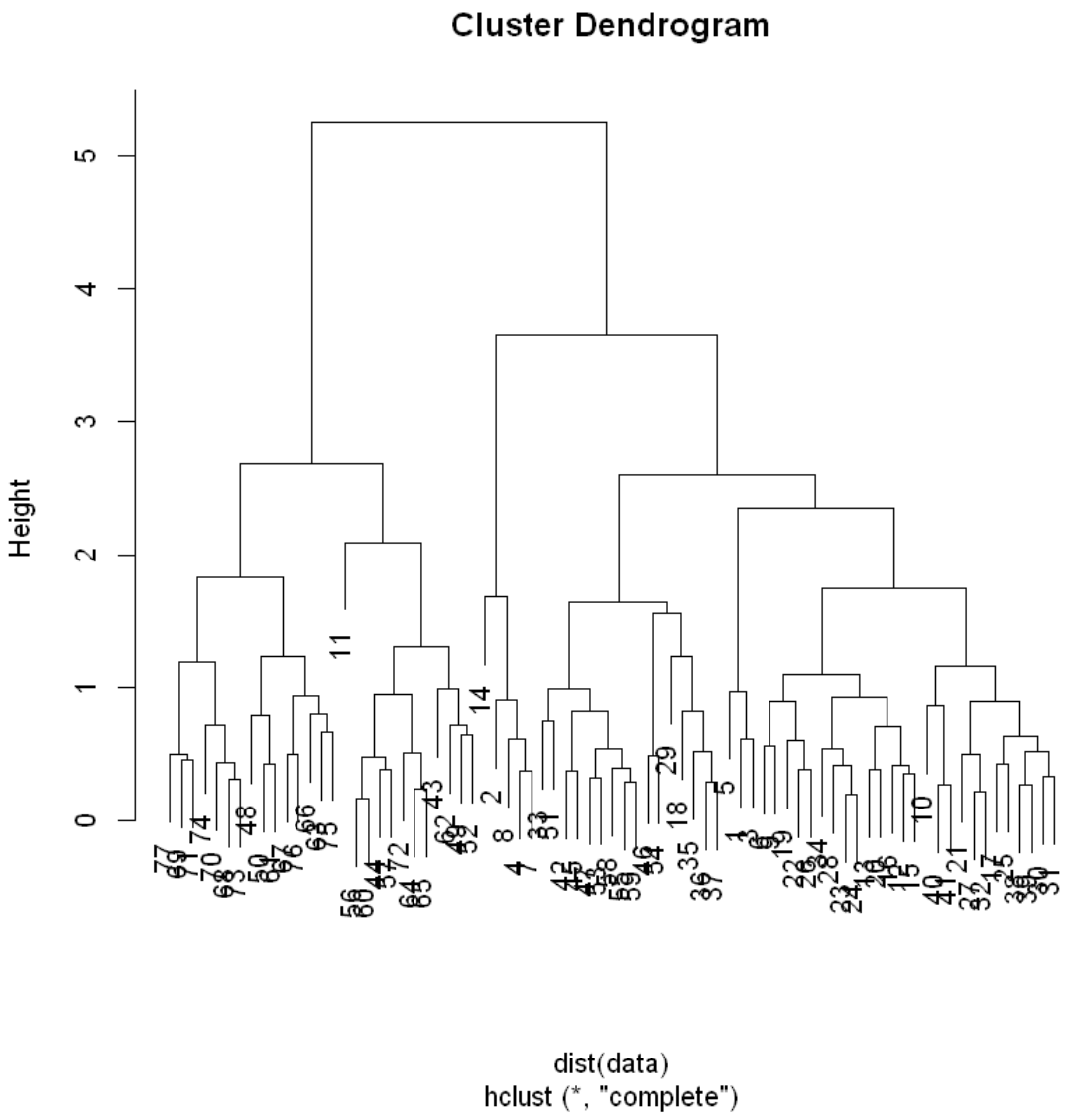
Чтож, теперь распределения достаточно симметричны. Есть два выброса в CallsCount (78, 79). Удалим их.

In [7]:
```
data[78,]<-NA
data[79,]<-NA
data<-na.omit(data)
```

3. Постройте кластеризацию методом k-средних. Попробуйте построить несколько решений и выберите самое, на ваш взгляд, лучшее, основываясь на любом из методов, который мы разбирали в курсе. Опишите в работе, как вы отбирали оптимальную модель.

Для начала попробуем определить оптимальное число кластеров на основе дендрограммы иерархического кластерного анализа.

In [8]:
```
plot(hclust(dist(data)))
```



Визуальный анализ показывает, что разбиение на 6-7 кластеров будет достаточно информативно. Создадим несколько разбиений и сравним их по эвристическому критерию - наполненности кластеров.

In [9]:
```
k5<-kmeans(data, centers=5, nstart=50)
k6<-kmeans(data, centers=6, nstart=50)
k7<-kmeans(data, centers=7, nstart=50)
k8<-kmeans(data, centers=8, nstart=50)
```

In [10]:
```
k5$size
k6$size
k7$size
k8$size
```

```
20  20  17  11  9

9  19  11  4  17  17

5  17  11  4  9  19  12

12  11  16  9  5  8  12  4
```

In [11]:
```
k5$centers
```

AddressCount	CallsCount	ClicksCount	FirmsCount	GeoPart	MobilePart	UsersCount	Distance
1.622576	2.362358	3.068272	1.723982	0.2840490	0.3427446	2.748341	3.559526
2.349122	2.968571	3.682081	2.050177	0.3114568	0.4142210	3.337550	3.444730
3.260037	3.475750	4.598880	2.679454	0.4168509	0.5957511	4.294168	3.159411
3.238794	3.758394	4.560905	2.499151	0.3318177	0.4282027	4.156288	3.421094
2.418267	2.455658	3.625530	2.251119	0.3958638	0.4573768	3.387838	3.107804

С точки зрения эвристического критерия полноты кластеров (не менее 10% объектов) - примем в качестве пригодного решение с 5 кластерами.

4. Оцените, все ли признаки существенно различают группы, не нужно ли исключить какой-либо из них. Если исключение необходимо, укажите и объясните это в работе.

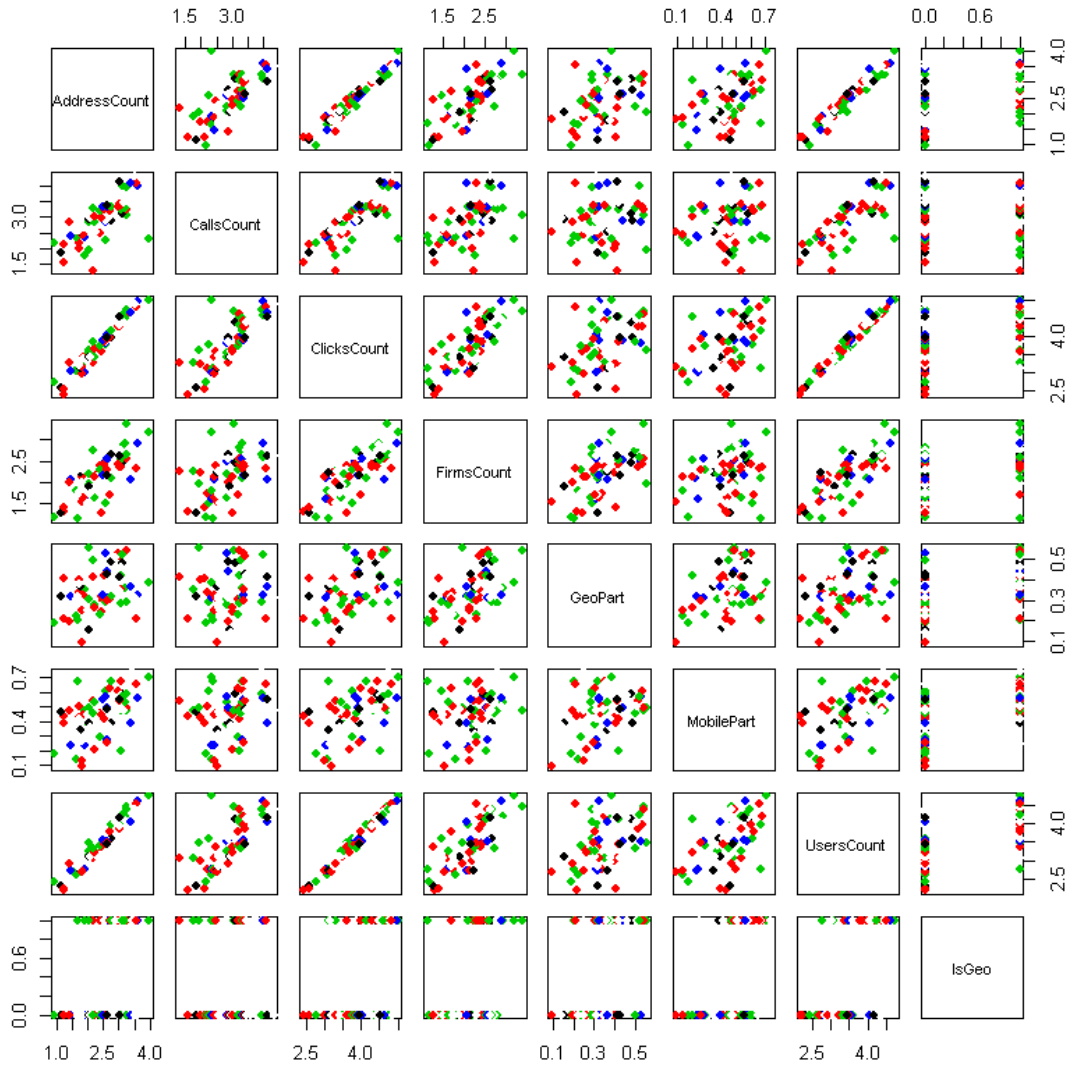
С точки зрения разброса расстояний центров возможно стоит исключить из рассмотрения признак Distance. Возможно также нет достаточно значимых различий между кластерами по GeoPart и MobilePart, однако исключать мы их не будем ввиду ценности для интерпретации результатов.

In [12]:
```
data<-data[,c(1:7,9)]
k5<-kmeans(data, centers=5, nstart=50)
k5$centers
data$cluster<-k5$cluster
```

AddressCount	CallsCount	ClicksCount	FirmsCount	GeoPart	MobilePart	UsersCount	IsGeo
2.418267	2.455658	3.625530	2.251119	0.3958638	0.4573768	3.387838	1.00
2.349122	2.968571	3.682081	2.050177	0.3114568	0.4142210	3.337550	0.00
3.260037	3.475750	4.598880	2.679454	0.4168509	0.5957511	4.294168	1.00
1.622576	2.362358	3.068272	1.723982	0.2840490	0.3427446	2.748341	0.05
3.238794	3.758394	4.560905	2.499151	0.3318177	0.4282027	4.156288	0.00

5. Опишите кластеры содержательно: какие рубрики попали в какой из кластеров? В чём особенности кластеров? Можете придумать группам какие-нибудь яркие, запоминающиеся названия, которые отражают особенности построенных кластеров.

In [13]:
```
plot(data[1:8], col=k5$centers, pch=19, cex=1)
```



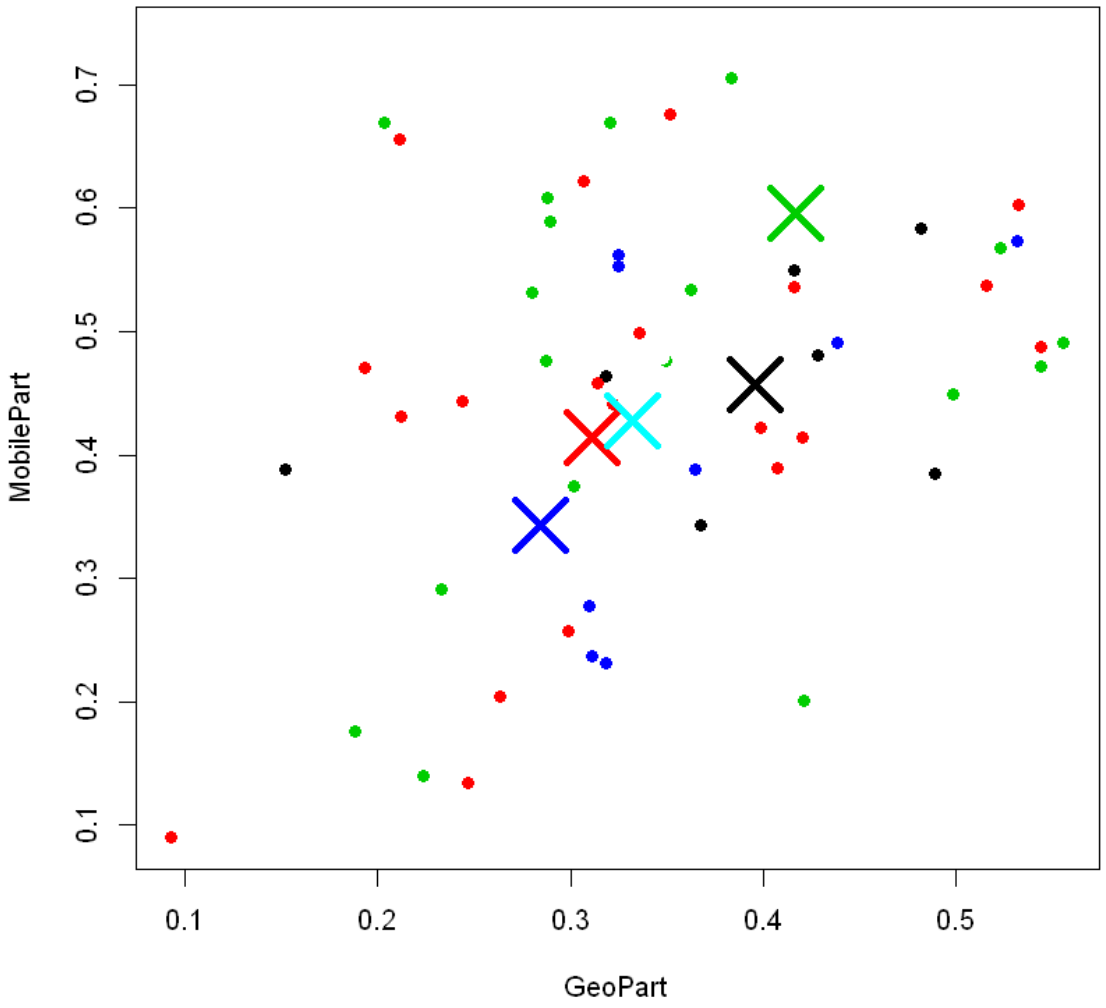
Можно заметить что в совокупности выделяются две пары групп, в среднем похожих, но отличающихся наличием или отсутствием геозависимости (две группы IsGeo=1 и две группы IsGeo=0). Также выделяется одна принципиально отличающаяся по всем признакам группа с центром 0.05 по IsGeo.

В каждой паре (геозависимой и геонезависимой) выделяется группа с высокими значениями всех признаков и группа с низкими значениями. Исключительная группа (0.05 по IsGeo) имеет очевидно минимальные координаты центра по всем измерениям.

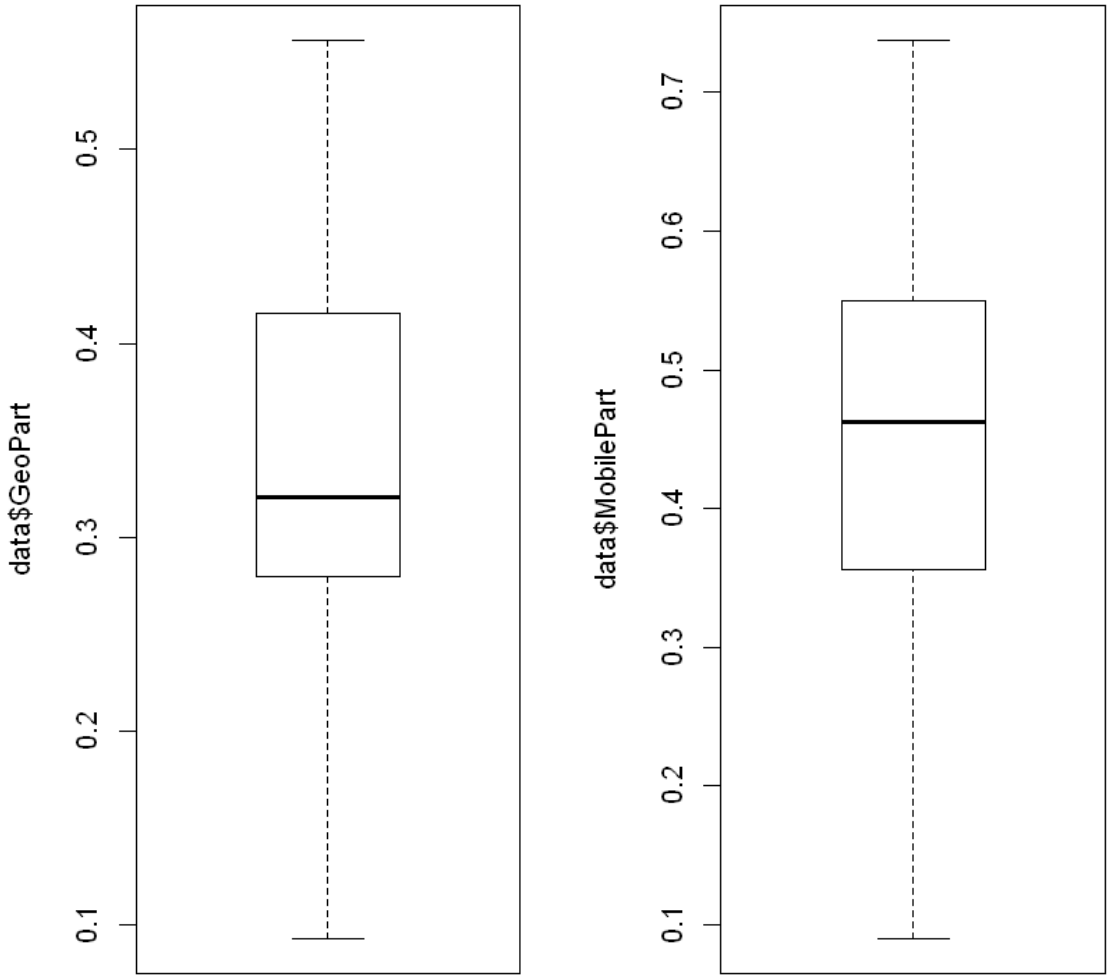
6. Сравните кластеры по доле трафика с карты и с мобильных продуктов (признаки GeoPart и MobilePart соответственно). Предварительно проанализируйте признаки, по которым будете сравнивать, и выберите соответствующий критерий для сравнения. Опишите выбор критерия: как выбирали, и почему остановились именно на этом?

Посмотрим для начала диаграмму рассеяния по данным признакам в кластерах с нанесением центров кластеров.

In [14]:
```
plot(data[,5:6], col=k5$centers, pch=19, cex=1)
points(k5$centers[,5:6], pch=4, cex=4, lwd=4, col=as.integer(rownames(k5$centers
[,5:6])))
```



In [15]:
```
par(mfcol=c(1,2), mar=c(4,4,2,2))
Boxplot(data$GeoPart)
Boxplot(data$MobilePart)
```



Данные признаки распределены достаточно симметрично и не обнаруживают выбросов. Проверим их нормальность используя достаточно мощный критерий Шапиро-Уилка. Заодно посмотрим Q-Q графики сравнения с нормальным распределением.

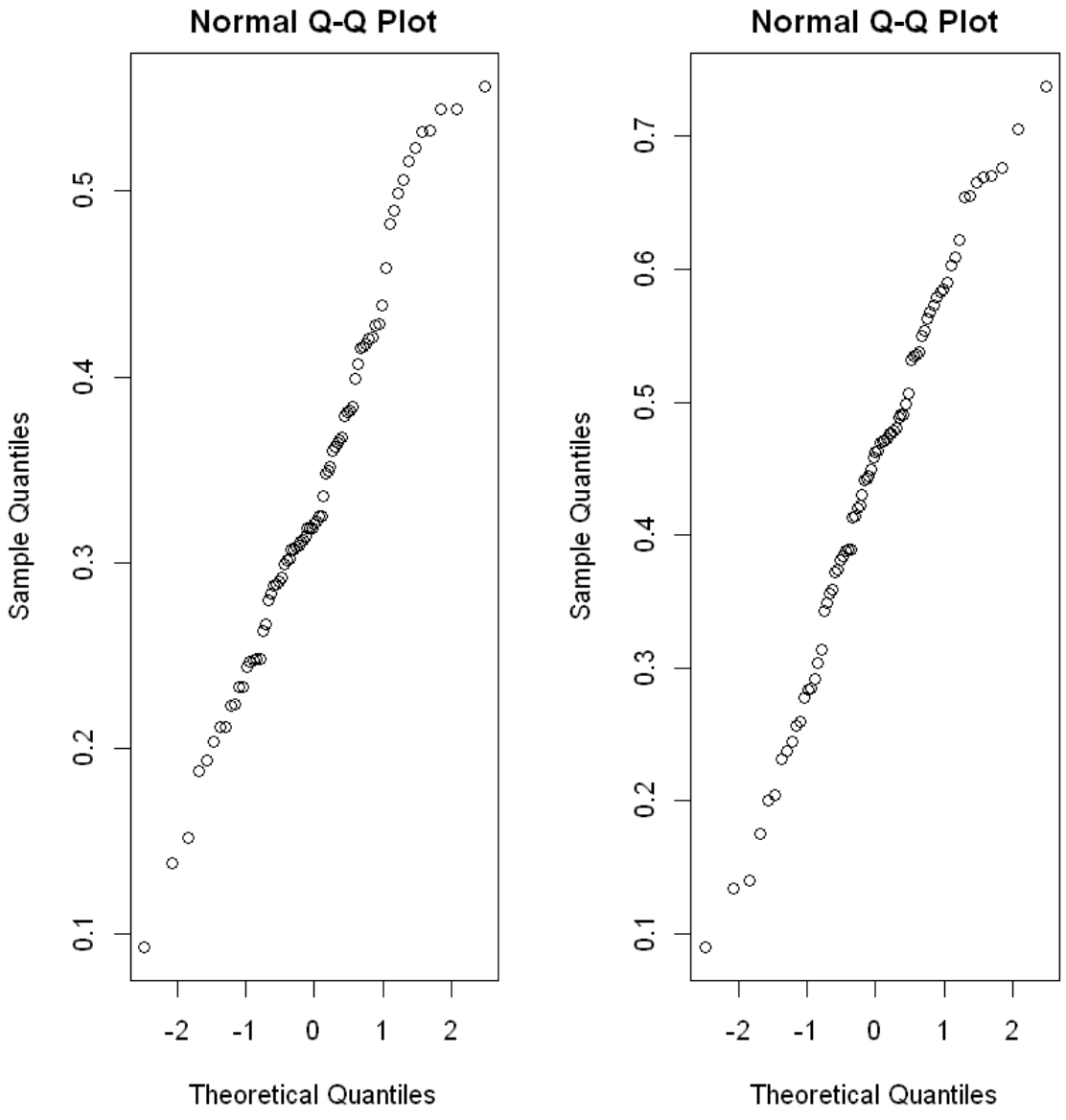
In [16]:
```
shapiro.test(data$GeoPart)
shapiro.test(data$MobilePart)
par(mfcol=c(1,2), mar=c(4,4,2,2))
qqnorm(data$GeoPart)
qqnorm(data$MobilePart)
```

```
        Shapiro-Wilk normality test

data:  data$GeoPart
W = 0.97796, p-value = 0.2014

        Shapiro-Wilk normality test

data:  data$MobilePart
W = 0.98441, p-value = 0.4683
```



p-value критерия Шапиро-Уилка значительно превышает 0.05, поэтому данные в целом можно признать распределенными нормально. Это значит, что мы можем использовать критерии сравнения кластеров (как несвязанных выборок) основанные на распределениях стьюдента и Фишера, в том числе ANOVA.

7. Сделайте содержательный вывод о различиях между кластерами, а также вывод о статистической значимости выявленных различий. Различаются ли кластеры (группы рубрик/сфер) по доле трафика с карты и с мобильных продуктов? Как именно? Значимы ли эти различия статистически?

Воспользуемся возможностью выполнить дисперсионный анализ (ANOVA).

In [17]:
```
with(data, summary(aov(GeoPart ~ cluster))) #ANOVA
with(data, summary(aov(MobilePart ~ cluster)))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
cluster      1 0.0251 0.02514   2.367  0.128
Residuals   75 0.7966 0.01062

            Df Sum Sq Mean Sq F value Pr(>F)
cluster      1 0.0338 0.03383   1.583  0.212
Residuals   75 1.6025 0.02137
```

Результаты дисперсионного анализа показывают, что в случае GeoPart и MobilePart кластеры статистически значимо различаются на уровне p>0.05 (имеются как минимум два кластера, отличающихся в среднем).

In [ ]:
```
```