Практическая часть

25.01.2023

```
library(ggplot2)
library(tidyverse)
library(geojsonio)
```

Карта районов

```
link = 'https://raw.githubusercontent.com/brianzelip/which-baltimore-neighborhood/master/data/Neighborhoods.geojson'
spdf <- geojson_read(link, what = "sp")
fortified <- fortify(spdf, region = "name")

ggplot() + geom_polygon(data = fortified, aes(x = long, y = lat, group = group), fill = "white", color = "grey") + theme_void() + coord_map()
```

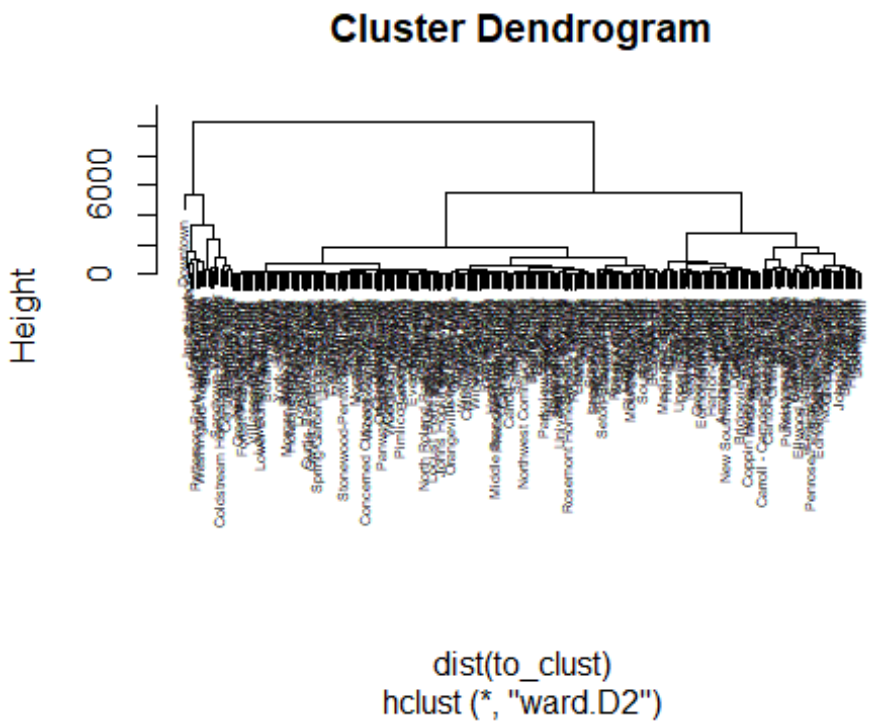
Загрузка датафрейма и отбор переменных:

```
crime = read.csv(file.choose())

to_clust <- crime %>% select(ASSAULT, BURGLARY, HOMICIDE, LARCENY, RAPE, ROBBERY)
rownames(to_clust) <- crime$Neighborhood

library(factoextra)
library(NbClust)
library(fossil)
```

Иерархический кластерный анализ

Будет использовано подходящее для метода Уорда Евклидово расстояние. Метод Уорда выбран как самый эффективный.

```
hc <- hclust(dist(to_clust), method = 'ward.D2')
plot(hc, cex=0.45)
```




```
ward <- cutree(hc, k = 3)
to_clust$cluster <- factor(ward)
crime$ward <- factor(ward)
```

Оценка качества кластеризации

Описательные статистики

```
summary(subset(to_clust, ward==1))
```

```
##     ASSAULT         BURGLARY         HOMICIDE         LARCENY
##  Min.   :  0.0   Min.   :  0.00   Min.   : 0.000   Min.   :  0.00
##  1st Qu.: 39.5   1st Qu.: 30.00   1st Qu.: 0.000   1st Qu.: 81.75
##  Median : 95.0   Median : 63.50   Median : 1.000   Median :144.50
##  Mean   :109.1   Mean   : 79.18   Mean   : 2.148   Mean   :156.55
##  3rd Qu.:166.2   3rd Qu.:118.00   3rd Qu.: 3.000   3rd Qu.:205.75
##  Max.   :317.0   Max.   :317.00   Max.   :15.000   Max.   :482.00
##       RAPE           ROBBERY       cluster
##  Min.   : 0.000   Min.   :  1.00   1:176
##  1st Qu.: 1.000   1st Qu.: 16.00   2:  0
##  Median : 2.000   Median : 34.00   3:  0
##  Mean   : 2.761   Mean   : 38.09
```

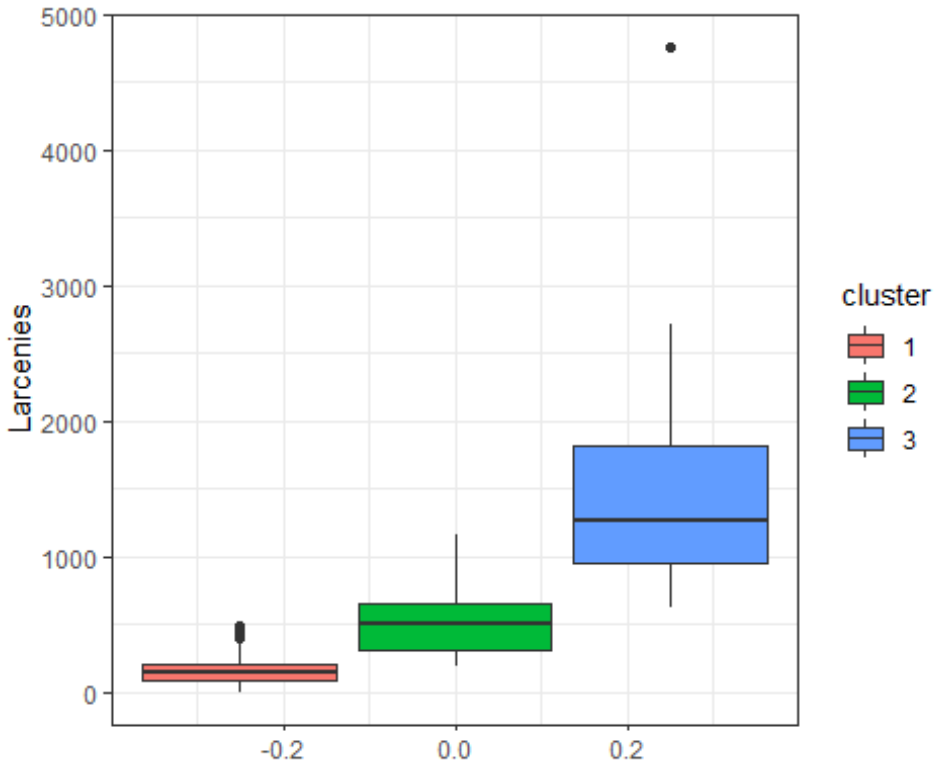
```
##  3rd Qu.: 4.000   3rd Qu.: 53.00
##  Max.   :15.000   Max.   :151.00

summary(subset(to_clust, ward==2))

##     ASSAULT         BURGLARY        HOMICIDE         LARCENY
##  Min.   : 64.0   Min.   : 22.0   Min.   : 0.000   Min.   : 189.0
##  1st Qu.:311.8   1st Qu.:154.5   1st Qu.: 3.000   1st Qu.: 302.0
##  Median :396.5   Median :203.0   Median : 7.000   Median : 503.0
##  Mean   :437.5   Mean   :232.4   Mean   : 8.329   Mean   : 519.2
##  3rd Qu.:506.5   3rd Qu.:290.8   3rd Qu.:12.000   3rd Qu.: 655.8
##  Max.   :948.0   Max.   :573.0   Max.   :31.000   Max.   :1165.0
##       RAPE           ROBBERY       cluster
##  Min.   : 1.000   Min.   : 33.00   1: 0
##  1st Qu.: 5.250   1st Qu.: 86.25   2:82
##  Median : 7.500   Median :122.50   3: 0
##  Mean   : 8.854   Mean   :133.09
##  3rd Qu.:11.000   3rd Qu.:171.75
##  Max.   :29.000   Max.   :277.00

summary(subset(to_clust, ward==3))

##     ASSAULT          BURGLARY         HOMICIDE        LARCENY
##  Min.   : 343.0   Min.   : 111.0   Min.   : 2.00   Min.   : 620.0
##  1st Qu.: 759.8   1st Qu.: 442.5   1st Qu.: 6.50   1st Qu.: 953.2
##  Median :1118.0   Median : 557.0   Median :19.50   Median :1264.0
##  Mean   :1197.0   Mean   : 590.4   Mean   :20.45   Mean   :1538.3
##  3rd Qu.:1559.5   3rd Qu.: 758.5   3rd Qu.:31.50   3rd Qu.:1821.5
##  Max.   :2892.0   Max.   :1235.0   Max.   :43.00   Max.   :4764.0
##       RAPE          ROBBERY       cluster
##  Min.   : 5.00   Min.   : 167.0   1: 0
##  1st Qu.:16.00   1st Qu.: 282.0   2: 0
##  Median :17.50   Median : 307.0   3:20
##  Mean   :24.80   Mean   : 372.2
##  3rd Qu.:31.75   3rd Qu.: 365.8
##  Max.   :83.00   Max.   :1146.0

to_clust %>% group_by(cluster) %>% summarise_at(vars(ASSAULT:ROBBERY),
.funs = mean)

## # A tibble: 3 × 7
##   cluster ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>     <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1          109.     79.2     2.15    157.  2.76    38.1
## 2 2          437.    232.      8.33    519.  8.85   133.
## 3 3         1197     590.     20.4    1538. 24.8    372.

to_clust %>% group_by(cluster) %>% summarise_at(vars(ASSAULT:ROBBERY),
.funs = median)

## # A tibble: 3 × 7
##   cluster ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>     <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1           95      63.5      1      144.   2       34
```

```
## 2 2          396.    203        7      503    7.5    122.
## 3 3         1118     557       19.5   1264   17.5    307
```

Как мы можем заметить, с переходом к более высокому (преступному) кластеру возрастает каждый показатель преступной активности, причём разница значительная. В первом кластере среднее значение количества нападений равно 109, во втором - уже 437, в третьем кластере 1197. Радует лишь одно: районов относящихся к первому, наименее преступному кластеру, больше количества двух других вместе взятых.

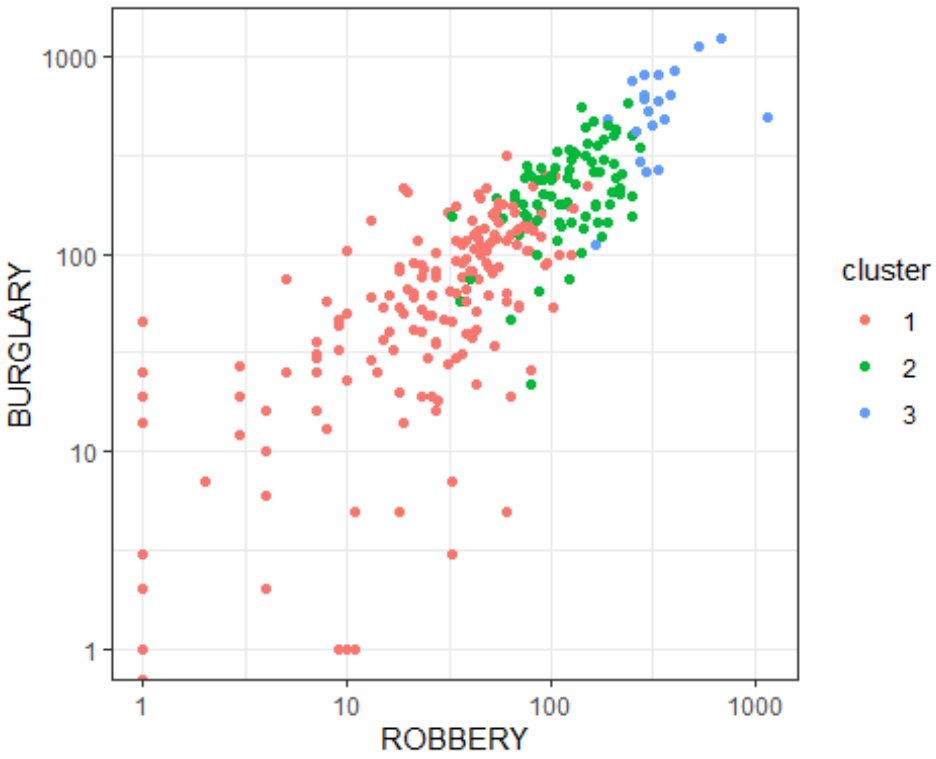
Визуализация

```
ggplot(data = to_clust, aes(y = LARCENY, fill = cluster)) +
  geom_boxplot() +
  theme_bw() +
  labs(y = "Larcenies")
```



На ящиках с усами мы можем увидеть сильные различия между кластерами по количеству хищений разного вида.

```
ggplot(data = to_clust, aes(x = ROBBERY, y = BURGLARY, color = cluster)) +
  geom_point() +
  theme_bw() +
  scale_x_log10() + scale_y_log10()

## Warning: Transformation introduced infinite values in continuous y-axis
```



На этом скаттерплоте построенными по количеству ограблений и краж со взломом, как резделены между собой кластеры. Однако некоторые наблюдения оказываются в окружении представителей другого кластера.

Статистические критерии

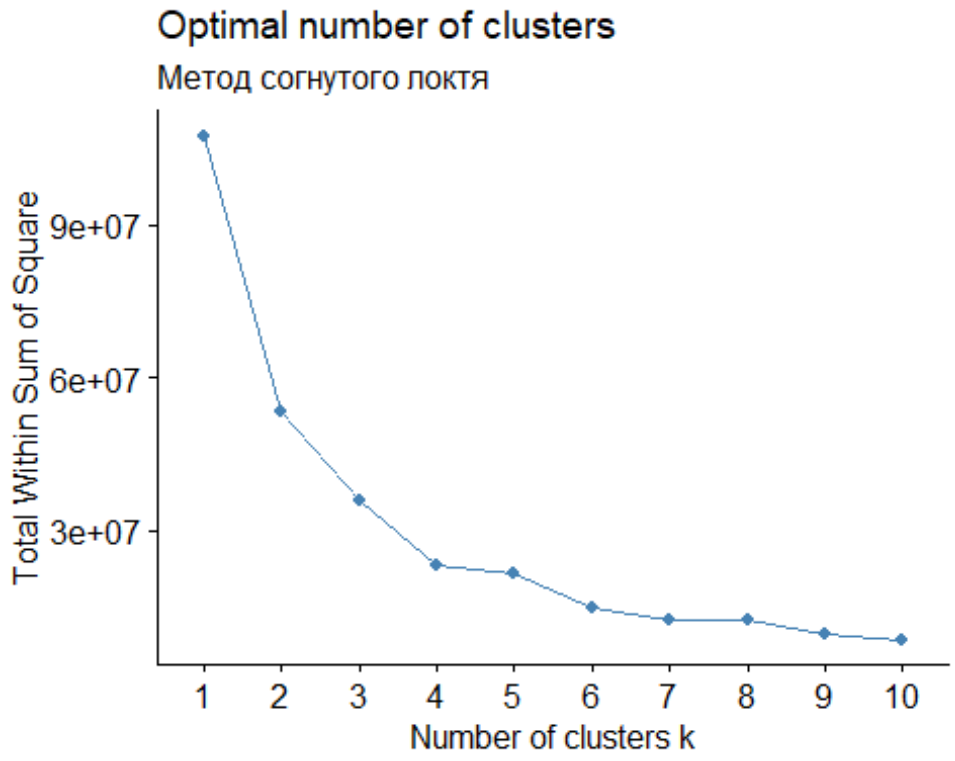
```
kruskal.test(crime$ASSAULT ~ crime$ward)

##
##  Kruskal-Wallis rank sum test
##
## data:  crime$ASSAULT by crime$ward
## Kruskal-Wallis chi-squared = 184.96, df = 2, p-value < 2.2e-16

kruskal.test(crime$BURGLARY ~ crime$ward)

##
##  Kruskal-Wallis rank sum test
##
## data:  crime$BURGLARY by crime$ward
## Kruskal-Wallis chi-squared = 143.36, df = 2, p-value < 2.2e-16

kruskal.test(crime$HOMICIDE ~ crime$ward)

##
##  Kruskal-Wallis rank sum test
##
## data:  crime$HOMICIDE by crime$ward
## Kruskal-Wallis chi-squared = 95.575, df = 2, p-value < 2.2e-16

kruskal.test(crime$LARCENY ~ crime$ward)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  crime$LARCENY by crime$ward
## Kruskal-Wallis chi-squared = 169.53, df = 2, p-value < 2.2e-16

kruskal.test(crime$RAPE ~ crime$ward)

##
##  Kruskal-Wallis rank sum test
##
## data:  crime$RAPE by crime$ward
## Kruskal-Wallis chi-squared = 136.73, df = 2, p-value < 2.2e-16

kruskal.test(crime$ROBBERY ~ crime$ward)

##
##  Kruskal-Wallis rank sum test
##
## data:  crime$ROBBERY by crime$ward
## Kruskal-Wallis chi-squared = 168.86, df = 2, p-value < 2.2e-16
```

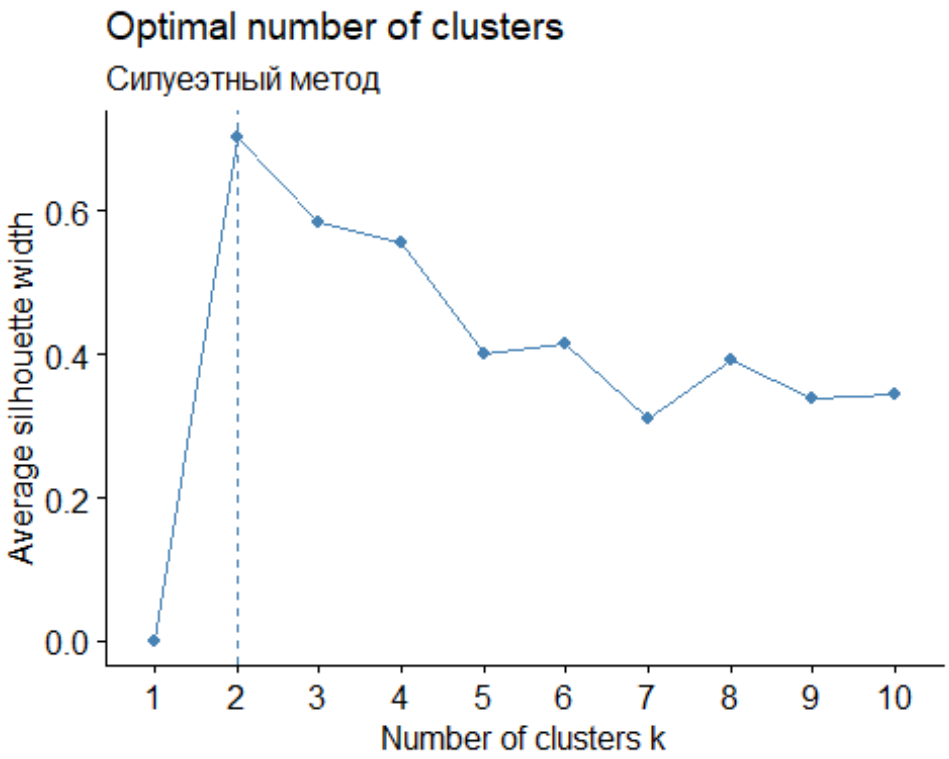
Все тесты подтверждают значимые различия между показателями различных кластеров.

Уточнение числа кластеров

```
fviz_nbclust(to_clust[1:6], kmeans, method = "wss") +
  labs(subtitle = "Метод согнутого локтя")
```



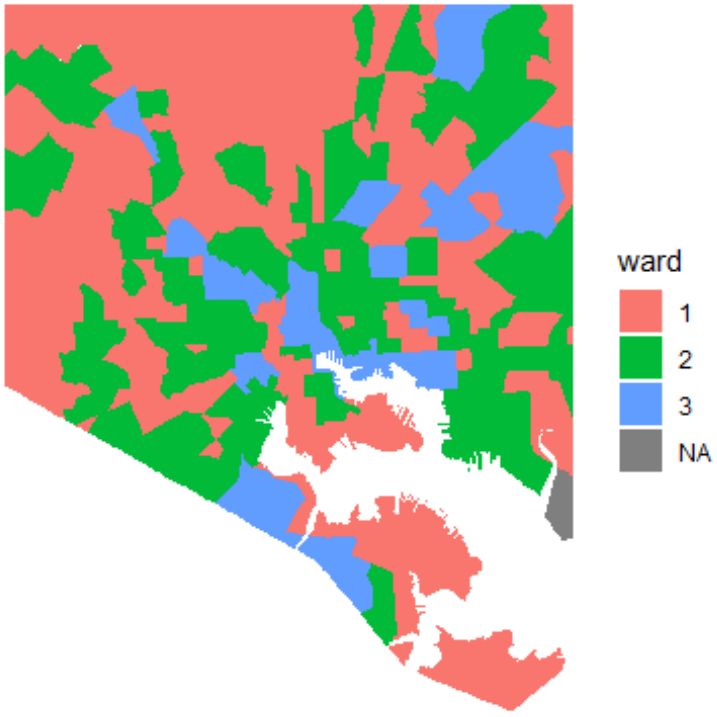
```
fviz_nbclust(to_clust[1:6], kmeans, method = "silhouette") +
  labs(subtitle = "Силуеэтный метод")
```



Предварительно выбранное количество кластеров кажется оптимальным. Метод согнутого локтя показывает, что с увеличением количества кластеров до четерыёх и дале результат улучшается намного меньше, чем при переходе от двух до трёх кластеров.

Кластеры и география

```
full <- fortified %>% left_join(. , crime, by=c("id"="Neighborhood"))

ggplot() + geom_polygon(data = full,
aes(fill = ward, x = long, y = lat, group = group)) +
theme_void() + coord_map()
```



На карте видна не очень отчётливая взаимосвязь следующего типа: районы центра города и близкие к центру относятся к третьему или второму кластерам, высоким по уровню преступности. По мере отдаления от центра города к окраинам становится больше районов относящимся к кластеру №1, наименее преступному.

K-means кластеризация

```
kclust <- kmeans(to_clust[1:6], 3)
crime$k <- factor(kclust$cluster)

crime %>% group_by(k) %>% summarise_at(vars(ASSAULT:ROBBERY),
.funs = mean)
```

```
## # A tibble: 3 × 7
##   k     ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>   <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1        562.    309.      9.84    729. 11.3    184.
## 2 2        148.     97.2     3.14    181.  3.41    46.5
## 3 3       1453.    614      22.7    1929. 31.7    437.
```

Кластеризация, проведённая методом kmeans, показывает немного отличающиеся от полученных выше результаты, тем не менее разделение примерно такое же, один кластер с низкой преступностью, второй со средней и третий с высокой, только в данном случае второй и первый кластер поменялись местами.

Вывод:

В первый кластер попали самые безопасные районы города, во втором носители средних показателей, в третьем кластере находятся районы, обладающие самыми высокими показателями преступности, причём при сравнений второго и третьего

кластера или первого и второго кластера, растут сразу все показатели преступности, без исключений

Для точного ответа на вопрос о влиянии географии на уровень преступности нужно больше данных. Мы не можем подтвердить или не подтвердить, например, гипотезы, о том что окраины менее преступны, из-за меньшего количества населения, количества ночных клубов, туристов и мигрантов. Тем более, что отсутствует чёткая взаимосвязь в виде понижения уровня преступности (согласно кластерам) по мере отдаления от центра города. На окраине также наблюдаются наиболее преступные районы, так и в центральных районах есть районы первого кластера, и они не единичны.