# hw1_msa_grinevich

polina grinevich

2023-01-29

# Подготовка среды

**Строим карту Балтимора:**

```
## Registered S3 method overwritten by 'geojsonsf':
##   method        from
##   print.geojson geojson
```

```
##
## Attaching package: 'geojsonio'
```

```
## The following object is masked from 'package:base':
##
##     pretty
```

```
spdf <- geojson_read('https://raw.githubusercontent.com/brianzelip/which-baltimore
-neighborhood/master/data/Neighborhoods.geojson',  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()
```

**Готовим данные по районам для кластеризации:**

```
crime <- read.csv('/Users/polinagrinevich/Downloads/BaltimoreCrimesAgg.csv')
```

# Отбор переменных

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ─
─
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.5.0
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ✔ purrr   1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ─
─
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

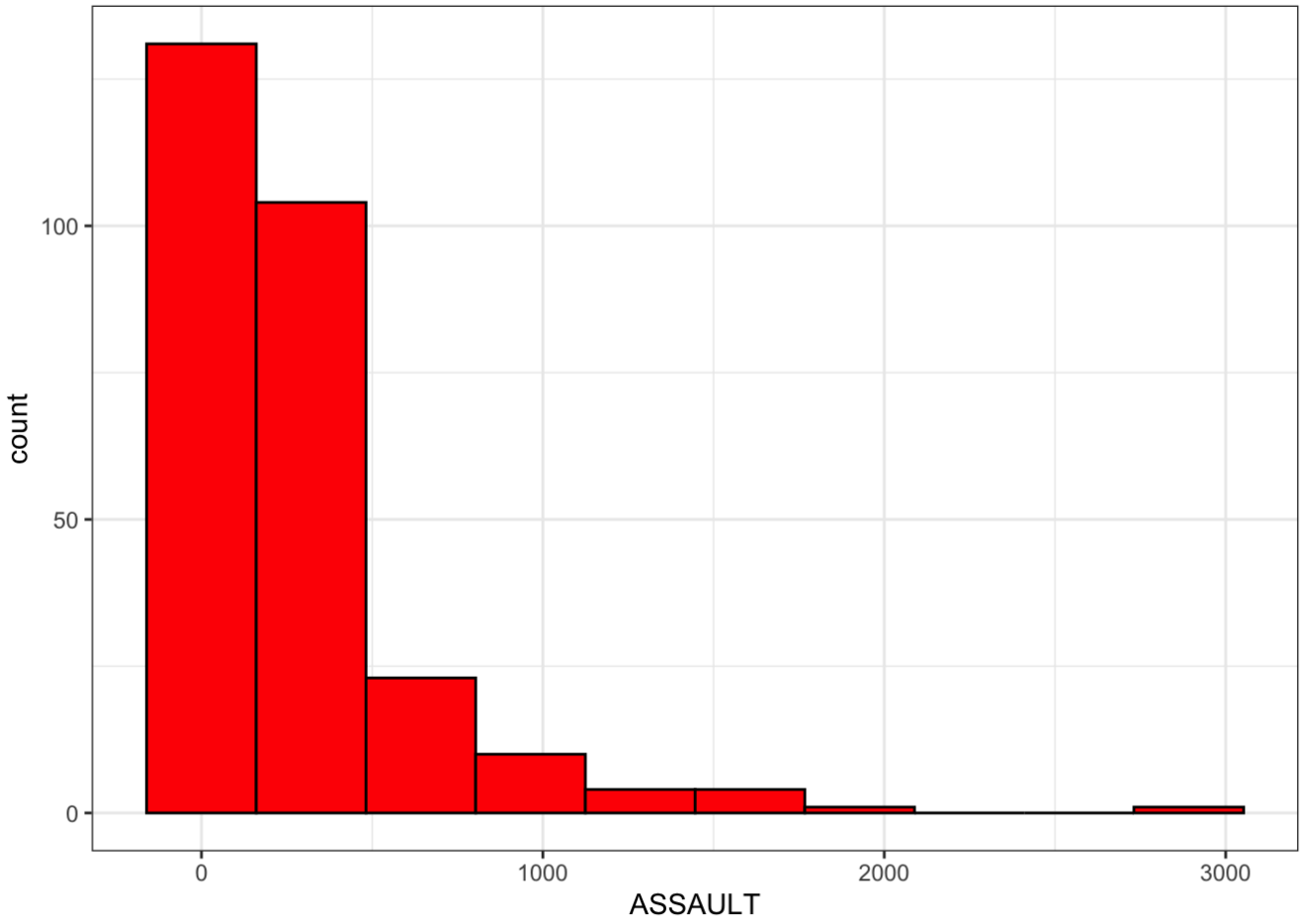
```
to_clust <- crime %>% select(-Neighborhood)
rownames(to_clust) <- crime$Neighborhood
to_clust <- na.omit(to_clust)
```

В датафрейм **to_clust** сохраняем все нужные нам столбцы из датасета crime, а именно все столбцы, кроме столбца с названиями районов. Строки датафрейма называем так же, как названия соответсующих районов. Удаляем пропущенные значения.

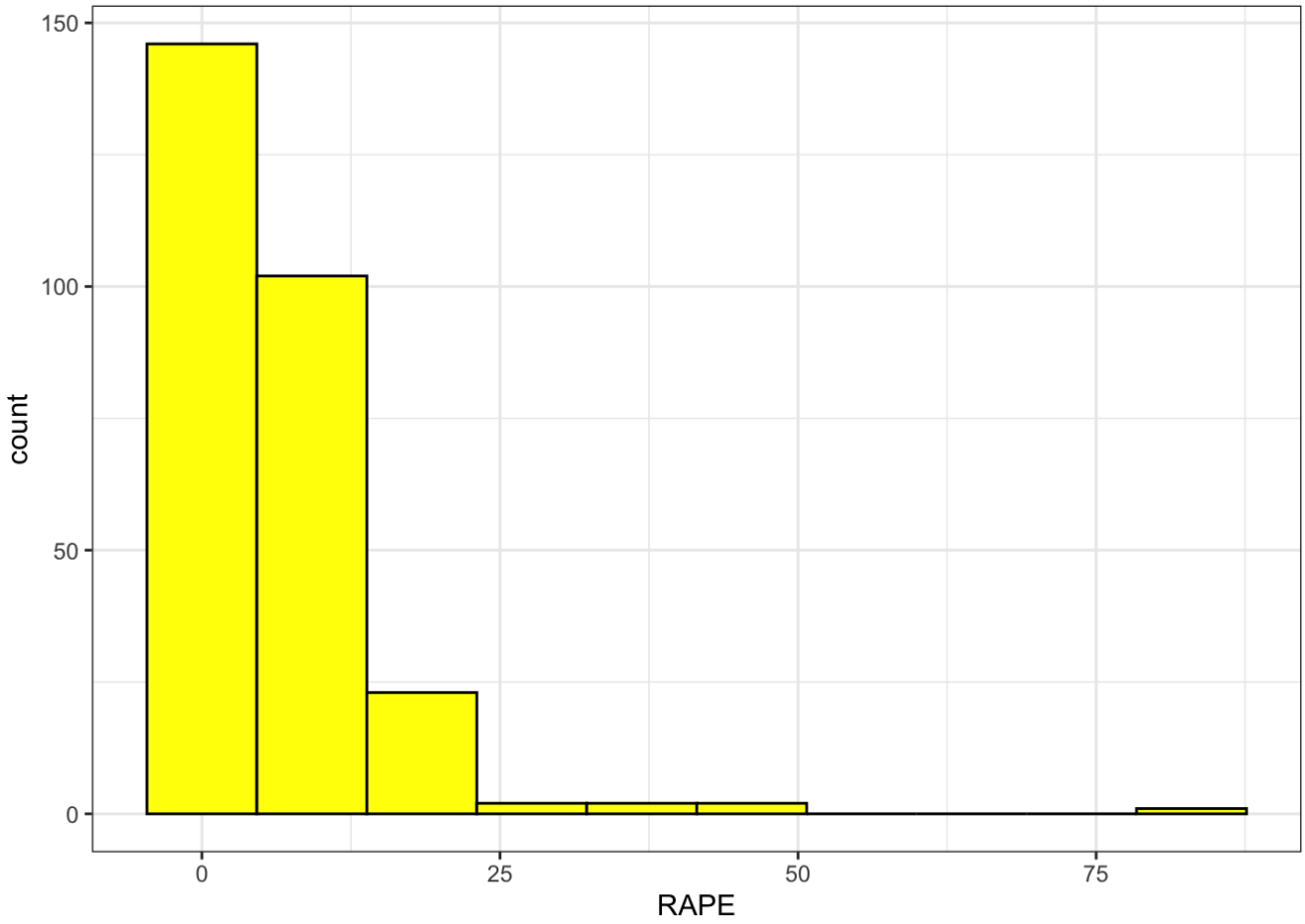
# Иерархический кластерный анализ

Прежде чем переходить к непосредственно иерархическому кластерному анализу, стоит обратиться к данным по каждой переменной и посмотреть, как они распределены в общем. Для этого я предлагаю построить гистограммы по каждому преступлению и посмотреть на ситуацию в общем, чтобы в дальнейшем, делая содержательный вывод по поводу того, сколько кластеров нам нужно, уже иметь представление о распределении преступлений.

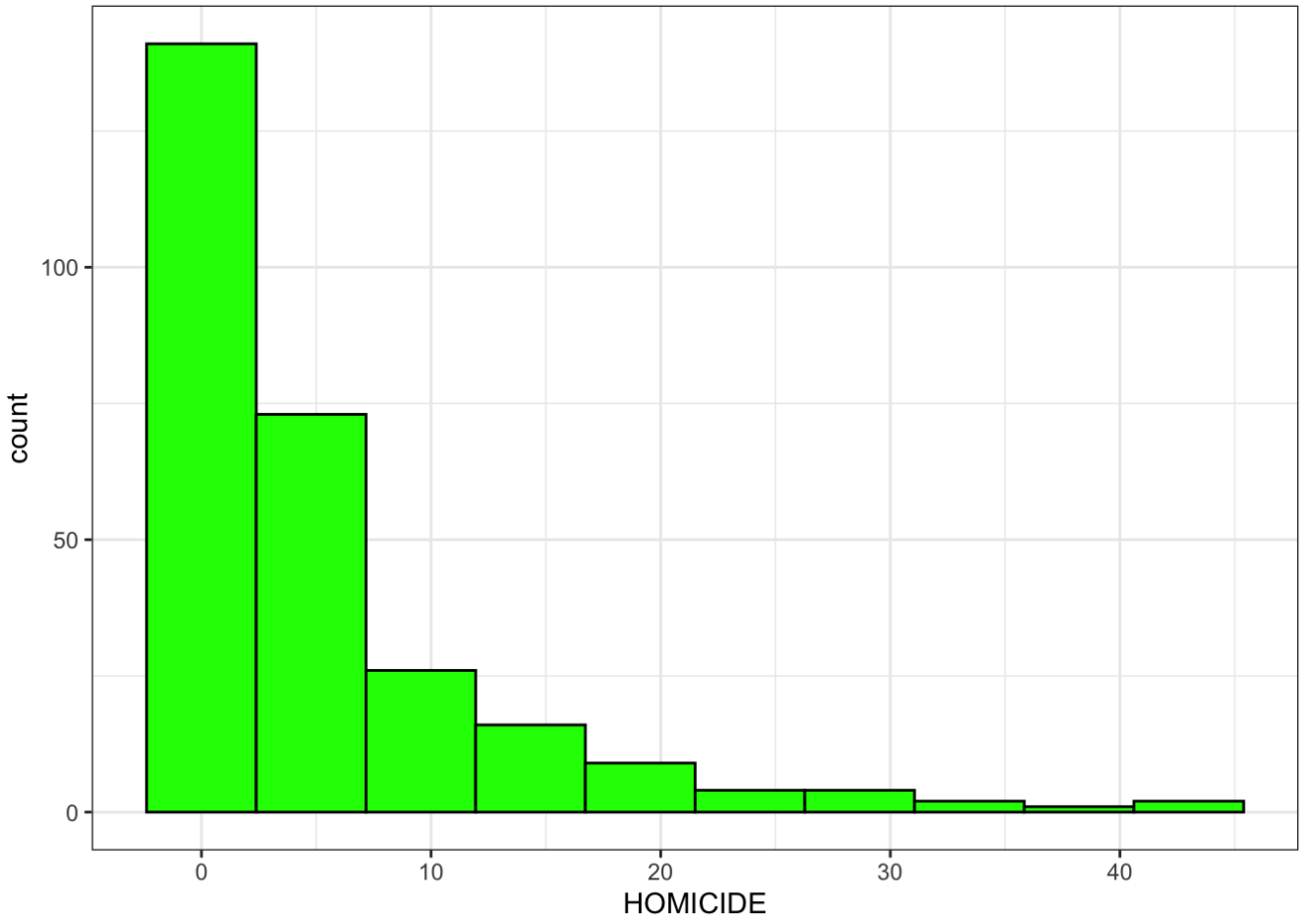
```
ggplot(data = to_clust, aes(x = ASSAULT)) +
  geom_histogram(fill = "red", color = "black", bins = 10) +
  theme_bw() +
  labs(x = "ASSAULT")
```



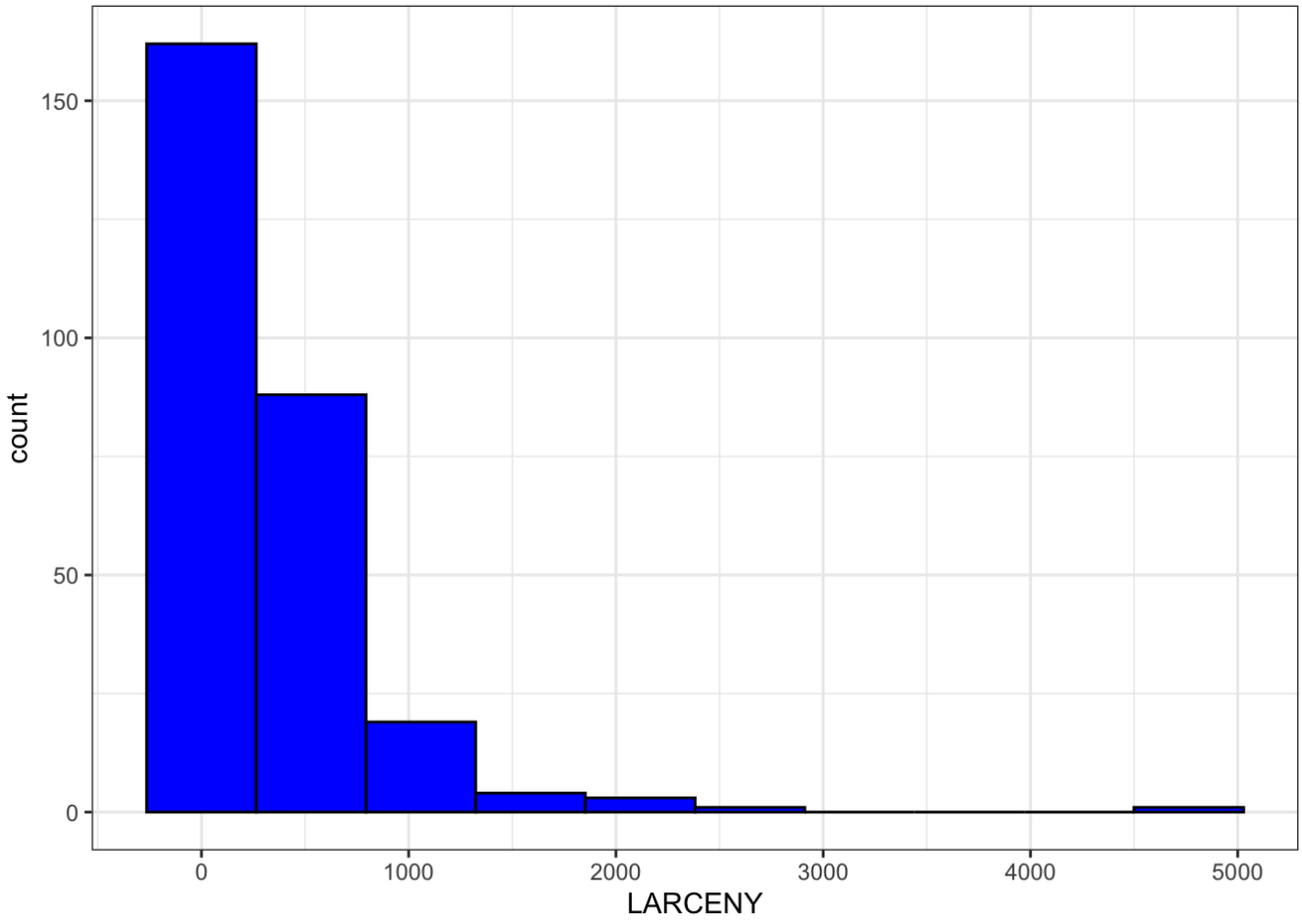

```
ggplot(data = to_clust, aes(x = RAPE)) +
  geom_histogram(fill = "yellow", color = "black", bins = 10) +
  theme_bw() +
  labs(x = "RAPE")
```



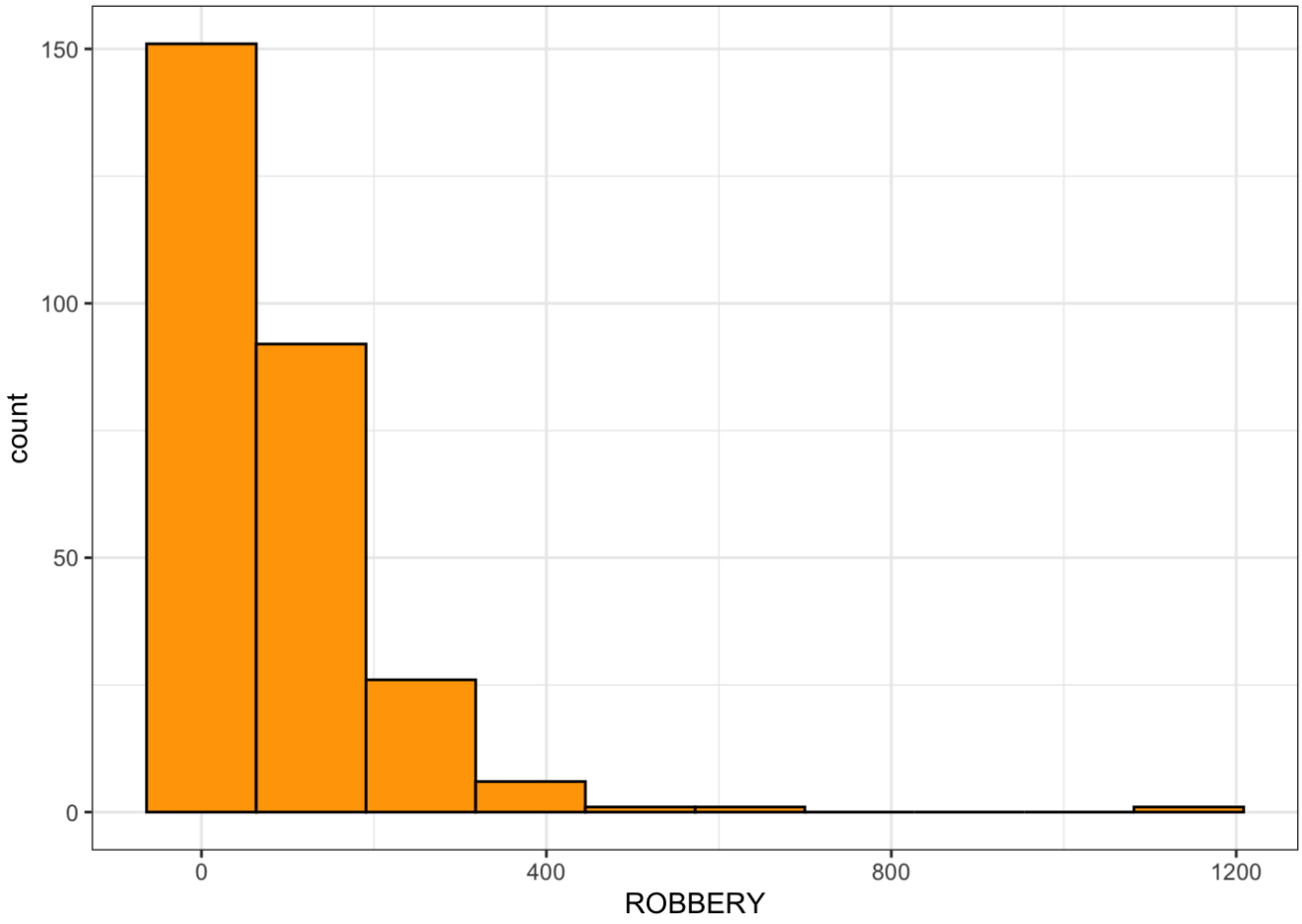

```
ggplot(data = to_clust, aes(x = HOMICIDE)) +
  geom_histogram(fill = "green", color = "black", bins = 10) +
  theme_bw() +
  labs(x = "HOMICIDE")
```




```
ggplot(data = to_clust, aes(x = LARCENY)) +
  geom_histogram(fill = "blue", color = "black", bins = 10) +
  theme_bw() +
  labs(x = "LARCENY")
```




```
ggplot(data = to_clust, aes(x = ROBBERY)) +
  geom_histogram(fill = "orange", color = "black", bins = 10) +
  theme_bw() +
  labs(x = "ROBBERY")
```




Распределение всех 6 видов преступлений выглядит похоже: много районов с отсутствием преступлений или с их малым количеством, затем идет повышение числа преступлений с понижением количества таких районов (чуть более опасные, но количество преступлений, кажется, находитсч в доупстимых рамках).

Затем идет небольшое количество районов с уже большим количеством преступлений. Несколько районов попадают в категорию достаточно опасных - это полученные очень низкие столбцы в центре каждой гистограммы.

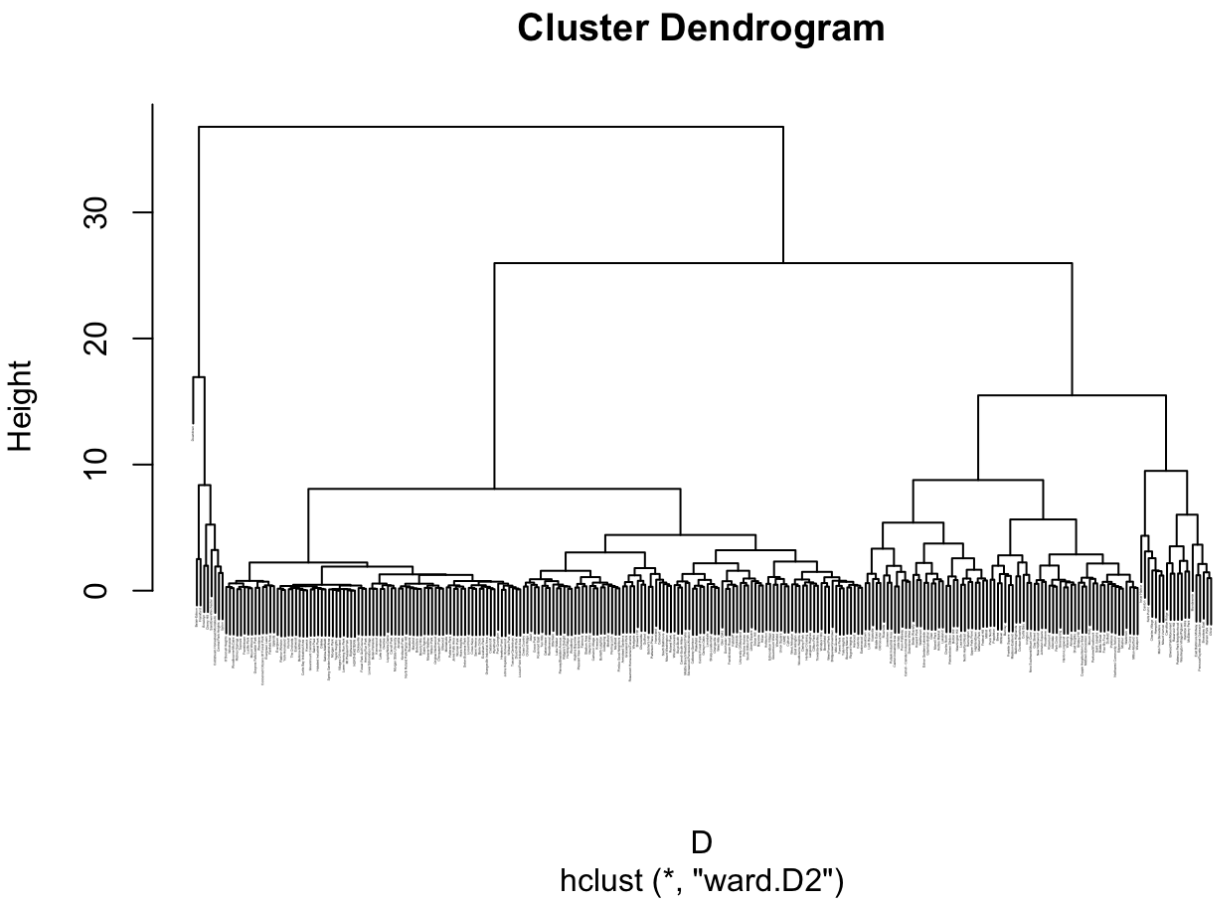
И *самое удивительное* - на каждой гистограмме видно, что существует какое-то маленькое количество районов с совсем аномально большим количеством преступлений - это крайние правые столбцы, которые по количеству преступлений обгоняют следующий по опасности район **сразу в несколько раз**.

# Реализция иерархического кластерного анализа

Метрика, которую мы будем использовать - **Евклидово расстояние**, а точнее его квадрат, потому что выбранный метод агрегирования - **метод Варда**, для которого используется квадрат Евклидового расстояния. Метод Варда основывается на присоединение точки к той группе, присоединение к которой приводит к наименьшему увеличению дисперсии в этйо группе, что делает этот метод наиболее статистически обоснованным. Евклидово расстоние хорошо работает с low-dimensional data, то есть с данными, у которых количество признаков по сравнению с количеством наблюдений незначительное. Этим я обосновываю выбор метрики и метода агрегирования.

Для метода Варда необходимо Евклидово расстояние в квдарате, однако встроенный в R метод ward.D2, который сам возводит значения расстояний в квадрат, этим мы и воспользуемся.

```
D <- dist(scale(to_clust))
hc <- hclust(D, method='ward.D2')
plot(hc, cex = 0.1)
```




Действительно, раннее предварительно выделенным кластером с аномально высоким уровнем преступности оказался **всего лишь один район**, который располагается слева. Он находится достаточно высоко: это означает, что его связь с ближайшим к нему кластером по значениям достаточно слаба. Такое аномальное значение не хочется относить к выбросам среди районов, а скорее к экстремальному случаю в одном из районов, на который стоит обратить внимание. Поэтому в наших интересах будет разрезать дендрогамму ниже значения ~16 для того, чтобы мы сделали кластер из одного наблюдения.

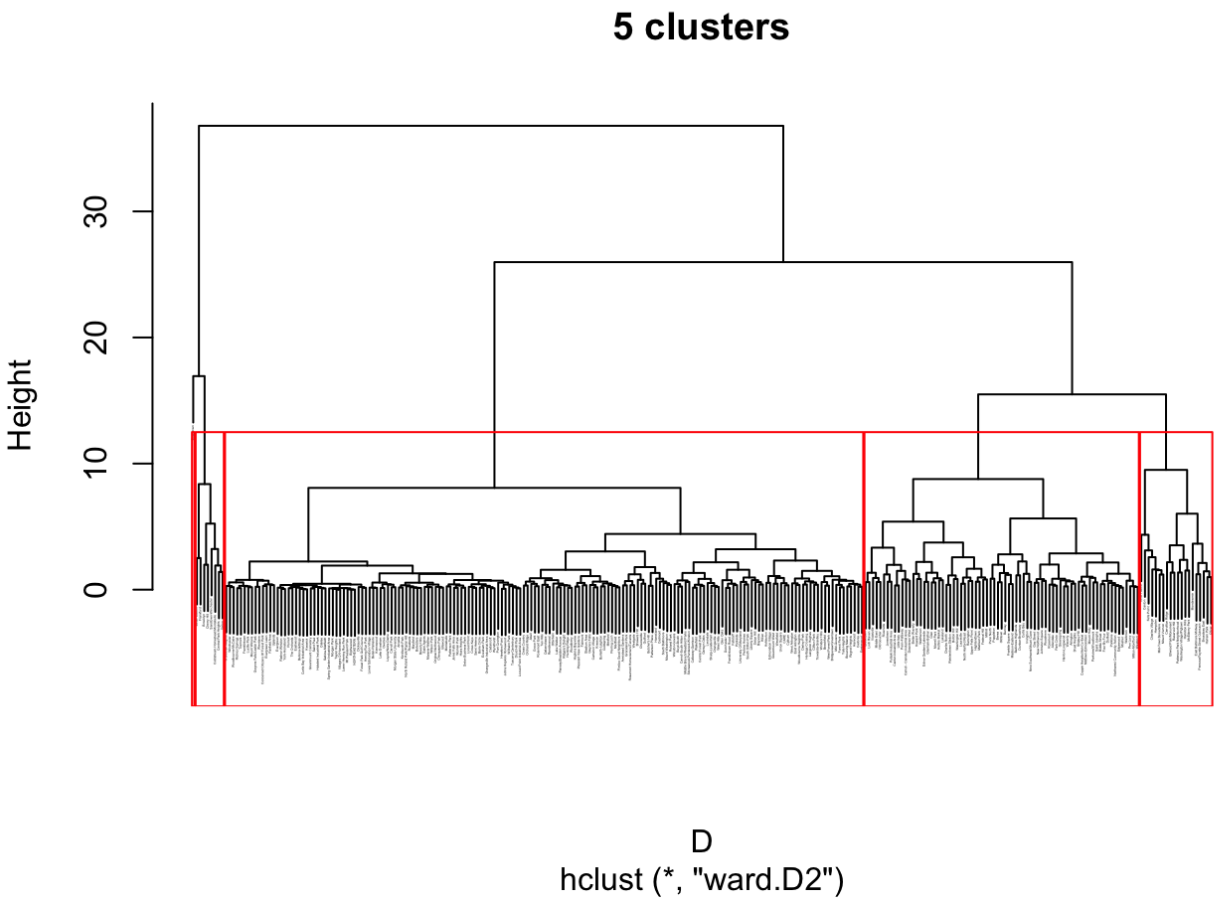
Мне представлется интересным разделить дендрогамму на **5 кластеров**, отражающих пять разных типов районов по степени преступности. Мы могли бы разделить на 4 кластера, и тогда примерно типы выглядели как *“аномально преступный - сильно преступный - преступный - спокойный”*, но такое деление 278 районов города всего лишь на 4 кластера мне как исследователю недостаточно. Кажется, что последние два кластера можно было бы разделить еще на 2, тогда категория “преступный” разделилась бы на 2 и учитывала бы районы, в которых есть районы с количеством преступлений в пределах нормы (и немного выше) и районы, которые можно уже более смело отнести к преступным.

Особенно важным такое деление кажется на фоне того, что районы города - это не однородные по своему строению деления города, где за одной чертой, проводимой на карте по этому району, сразу же занчительно снижается уровень преступности. Некоторые районы могут оказаться совсем новыми и полностью жилыми, в которых отсутствуют ночные клубы, места большого скопления людей (парки, торговые центры и т.д.), из-за чего уровень преступности в них значительно снижается, такая ситуация будет наблюдаться в кластере с самыми безопасными районами. Тогда следующий кластер мы относим к преступным, хотя внутри него могут быть существенные различия.

Если район внутри него окажется более старым и зажитым районом с развитой инфраструктурой, у которого и уровень преступнсоти будет выше из-за наличия третьих переменных (является ли он, например, экономическим или социальным центром города, есть ли там школы и т.д.), то он будет относиться к преступным, хотя на деле он не менее благополучный для простых жителей, ососбенно в сравнении с другими районами, не дотягивающими до уровня “сильно преступных”, но уже небезопасных. Мы разделим эти районы на 2 части, но будем иметь в виду, что незначительная преступность может объясняться не наличием в районе маньяков, мафии, якудзы или прочей неблагополучной прослойки общества, а иными проишествиями, которые все-таки нельзя оставлять без внимания.

Визуализируем 5 кластеров на дендрограмме:

```
plot(hc, cex = 0.1, main = "5 clusters")
rect.hclust(hc, k = 5, border="red")
```




Добавим в датафрейм to_clust столбец со значениями кластеров, не забудем сдлеать эти метки факторными:

```
cluster <- cutree(hc, k = 5)
to_clust$cluster <- factor(cluster)
```

# Оценка качества кластеризации

```
cluster01 <- to_clust %>% filter(cluster == 1)
cluster02 <- to_clust %>% filter(cluster == 2)
cluster03 <- to_clust %>% filter(cluster == 3)
```

```
cluster04 <- to_clust %>% filter(cluster == 4)
cluster05 <- to_clust %>% filter(cluster == 5)
```

В 1, 2 и 4 кластерах находятся много значений, поэтому на экран мы вывыдем лишь часть из них. 3 и 5 кластер содеражт по 8 и 1 районов соответсвенно, поэтому выведем все.

```
cluster01[1:5, 1:6]
```

```
##                   ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Abell                 147      122        1     269    1       90
## Arcadia               105       86        0     103    0       18
## Armistead Gardens     289      192        1     305    5       54
## Ashburton             115      249        6     230    1       94
## Barre Circle           28       49        1      69    0       25
```

```
cluster02[1:5, 1:6]
```

```
##                      ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Allendale                368      255       13     283    6       78
## Arlington                417      180       11     301    7      110
## Baltimore Highlands      729      199        7     677   17      216
## Barclay                  512      260       17     543   10      172
## Berea                    623      201       12     301   13       91
```

```
cluster03[, 1:6]
```

```
##                                ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Belair-Edison                    1759     1129       40    1919   31      530
## Brooklyn                         1594      852       30    1061   45      402
## Central Park Heights             1168      485       33     683   17      359
## Cherry Hill                      1548      806       20    1071   34      337
## Coldstream Homestead Montebello  1245      592       43     833   17      340
## Frankford                        1932     1235       35    1938   46      687
## Sandtown-Winchester              1750      522       42     945   36      298
## Upton                            1396      451       31    1099   23      316
```

```
cluster04[1:5, 1:6]
```

```
##                       ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Broadway East            1217      475       28     620   16      190
## Canton                    639      637        2    2174   10      289
## Carrollton Ridge          948      573       16     527   29      240
## Charles Village           328      346        2    1082   22      277
## East Baltimore Midway     859      284       31     862   12      203
```

```
cluster05[, 1:6]
```

```
##          ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Downtown    2892      486       19    4764   83    1146
```

Перед интерпреатцией данных было бы наглядно сразу вывести некоторые описательные статистики и описать кластеры не просто по абсолютным значениям преступлений в них, а в сравнении с некоторыми показателями (среднее, минимальное и максимальное значения) в других кластерах:

```
to_clust %>% group_by(cluster) %>% summarise_at(vars(ASSAULT:ROBBERY), mean)
```

```
## # A tibble: 5 × 7
##   cluster ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>     <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1          111.     80.8     1.79    161.  2.70    38.5
## 2 2          387.    206.      8.09    474.  7.88   120.
## 3 3         1549     759      34.2    1194. 31.1    409.
## 4 4          771.    438      13.0    1147. 15.8    247.
## 5 5         2892     486      19      4764  83     1146
```

```
to_clust %>% group_by(cluster) %>% summarise_at(vars(ASSAULT:ROBBERY), min)
```

```
## # A tibble: 5 × 7
##   cluster ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>     <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1             0        0        0       0     0       1
## 2 2            64       22        0     110     1      22
## 3 3          1168      451       20     683    17     298
## 4 4           328      111        0     465     5     163
## 5 5          2892      486       19    4764    83    1146
```

```
to_clust %>% group_by(cluster) %>% summarise_at(vars(ASSAULT:ROBBERY), max)
```

```
## # A tibble: 5 × 7
##   cluster ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>     <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1           380      317        8     482    12     129
## 2 2           755      555       22    1162    20     252
## 3 3          1932     1235       43    1938    46     687
## 4 4          1217      805       31    2712    29     386
## 5 5          2892      486       19    4764    83    1146
```

Первая табличка показывает нам среднее значение среди кластеров, вторая - минимальное, третье - максимальное по каждой переменной.

Итак: **что мы можем сказать по поводу полученных кластеров?**

# 1 кластер (174 района)

1 кластер является самым кластером с самыми **безопасными районами**. Средние значения всех видов преступлений меньше, чем во всех других районах, а минимальные значения по всем преступлениям, кроме ограблений, равны 0. Впрочем, минимальное значение ограблений равно 1.

Вкратце по поводу интрепретации таблицы: среднее значение нападений в первом кластере районов равно 110.75. Аналогичным образом интерпретируются остальыне показатели.

# 2 кластер (75 районов)

По степени безопасности 2 кластер районов становится на второе место: его средние значения выше, чем в первом кластере, но все еще ниже всех остальных по всем переменным.

# 4 кластер (20 районов)

3 кластер будет прокомментирован в следующем пункте для сохранения рейтинга районов по безопасности. 4 кластер собрал районы с уже заметным уровнем преступности: все средние значения преступлений выше чем в 1 и 2 кластерах, но ниже 3 и 5.

# 3 кластер (8 районов)

Небезопасные районы. Средние значение нападений 1549, краж со взломом 759, убийств 34, зищений 1146.6, изнасилований 31, ограбление 408. В этот раз не все из средних значений по 6 переменным ниже следующего по уровню преступности кластера: значение убийств в 3 кластере становится абсолютным максимум по всем кластерам. Эту странность можно будет объяснить попозже: нам необзодимо взглянуть на карту районов, разделенных на наши кластеры, и подумать, чем можно объяснить повышенное значений убийств.

# 5 кластер (1 район)

Тот самый район с аномальными значениями, замеченный еще на первых этапах обзора данных. 1 район с максимальными значениями преступлений по 5 из 6 переменными. Этим районом, на первый взгляд удивительно, оказался Downtown - центр города.

Хотя на самом деле большое количество престпулений в центре не так удивительно, если задуматься о том, почему статистика нам об этом говорит. Центр города концентрирует в себе большое количество магазинов всех ценовых сегментов, места большого скопления людей: достопримечательности, парки, большое количество заведений: от ресторанов до ночных клубов. В центре больше людей, у которых обычно есть что красть (например, туристы), а раз людей много - то еще и легче остаться незамеченным.

Этим может объясниться настолько большое количество грабежей и иных преступлений в центре. Факт того, что центр города является его самым опасным районом не феноменальный. Такое же явление наблюдается, например, и в Лос-Анджелесе, где центр города тоже стоит на первом месте по преступлениям. Объясняется это не только тем, что центр города - это одновременно и экономический и социальный центр, но и большим количеством бездомных и иммигрантов, которые чаще всего находятся именно в центре (https://propertyclub.nyc/article/most-dangerous-neighborhoods-in-los-angeles).

Прежде чем проводить дальнейший анализ, поменяем местами кластерами 3 и 4, чтобы числовое значение факторов соответствовало рейтинговому порядку от самого безопасного до самого опасного кластера:

```
cluster <- factor(cluster, levels = c(1, 2, 4, 3, 5))
to_clust$cluster <- cluster
```

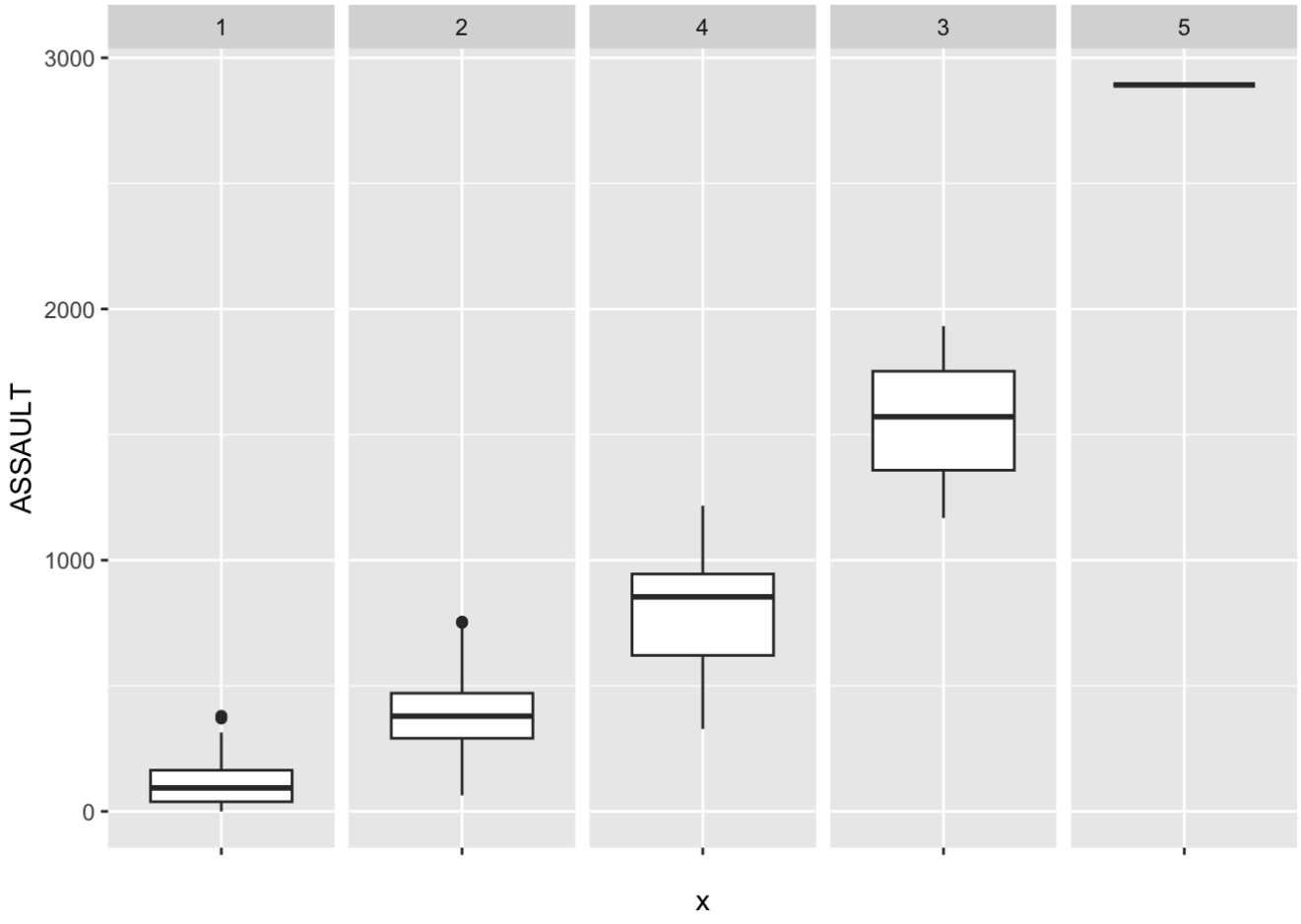
```
to_clust %>% group_by(cluster) %>% summarise_at(vars(ASSAULT:ROBBERY), mean)
```

```
## # A tibble: 5 × 7
##   cluster ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>     <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1          111.     80.8     1.79    161.  2.70    38.5
## 2 2          387.    206.      8.09    474.  7.88   120.
## 3 4          771.    438      13.0    1147. 15.8    247.
## 4 3         1549     759      34.2    1194. 31.1    409.
## 5 5         2892     486      19      4764  83     1146
```

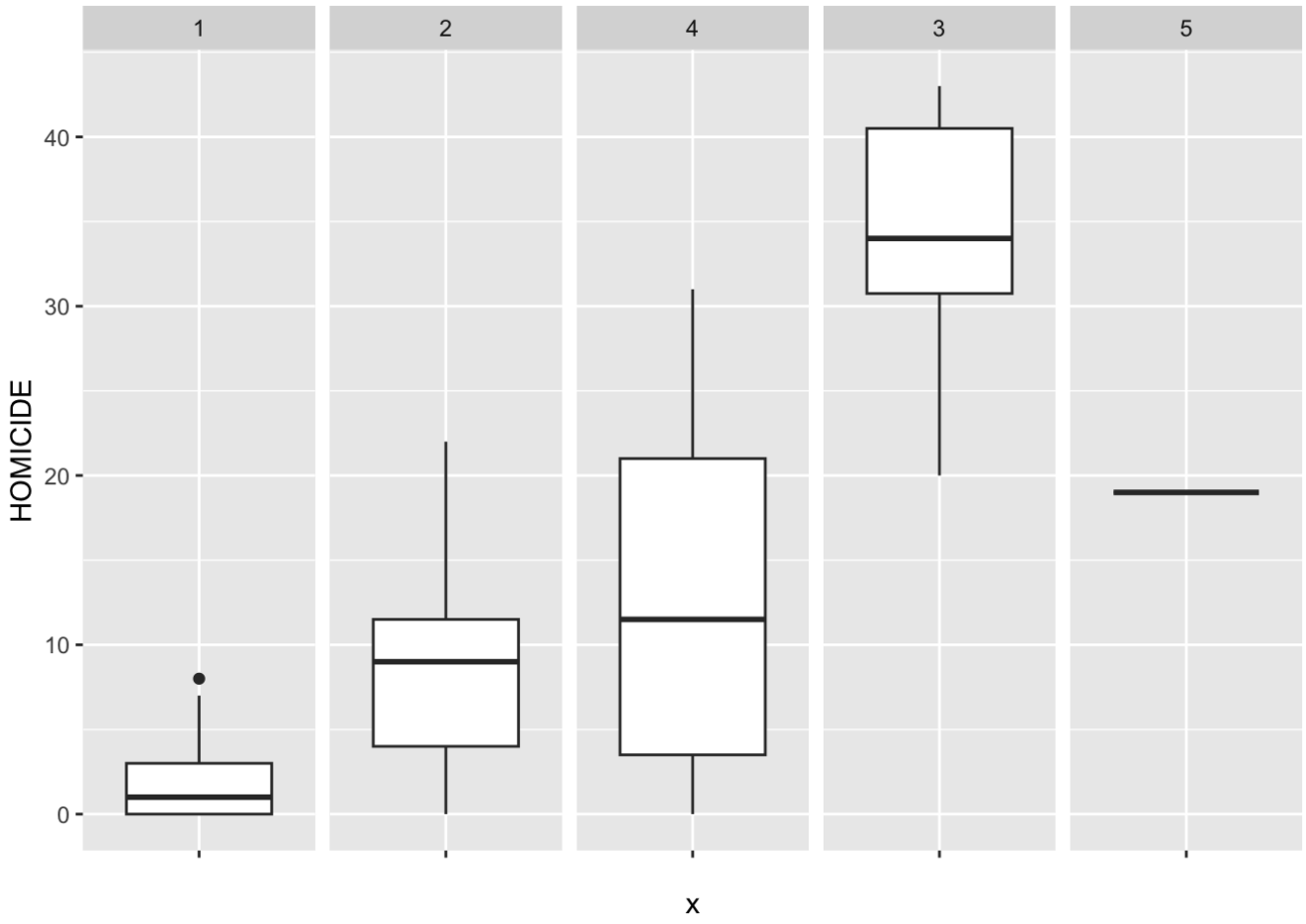
Проверяем, теперь всё на месте.

Посмотрим на распределение переменной, отвечающей за количество нападений, по кластерам. Поскольку в остальных переменных (за исключением переменной Убийства, в которой 3 кластер оказался выше, чем 5) наблюдается похожее распределение по преступлениям, мы не будем строить графики для всех. Визуализируем распределение с помощью ящиков с усами:

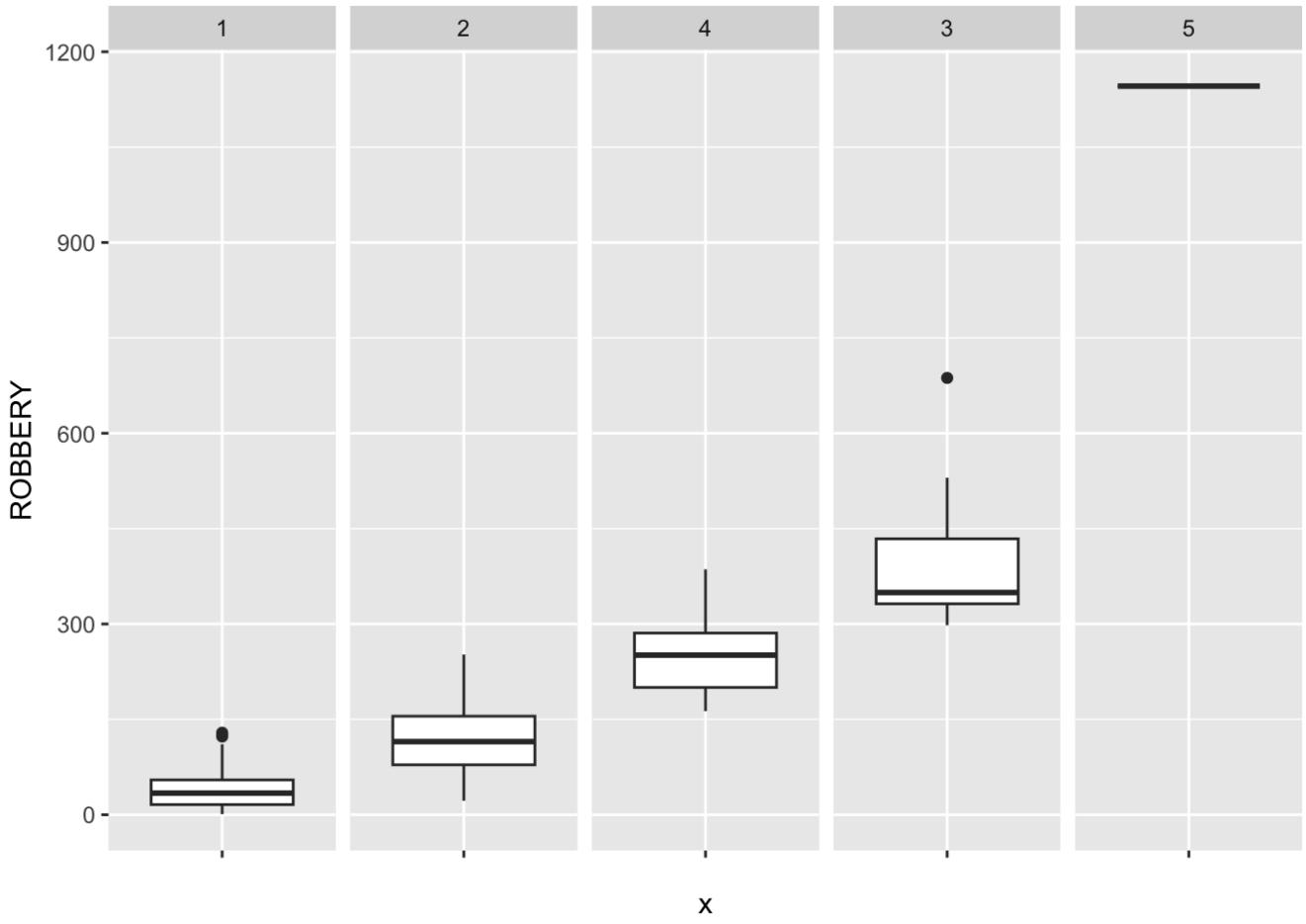
```
library(ggplot2)
ggplot(data = to_clust, aes(x = "", y = ASSAULT)) + geom_boxplot() + facet_grid(~c
luster)
```




```
ggplot(data = to_clust, aes(x = "", y = HOMICIDE)) + geom_boxplot() + facet_grid(~
cluster)
```

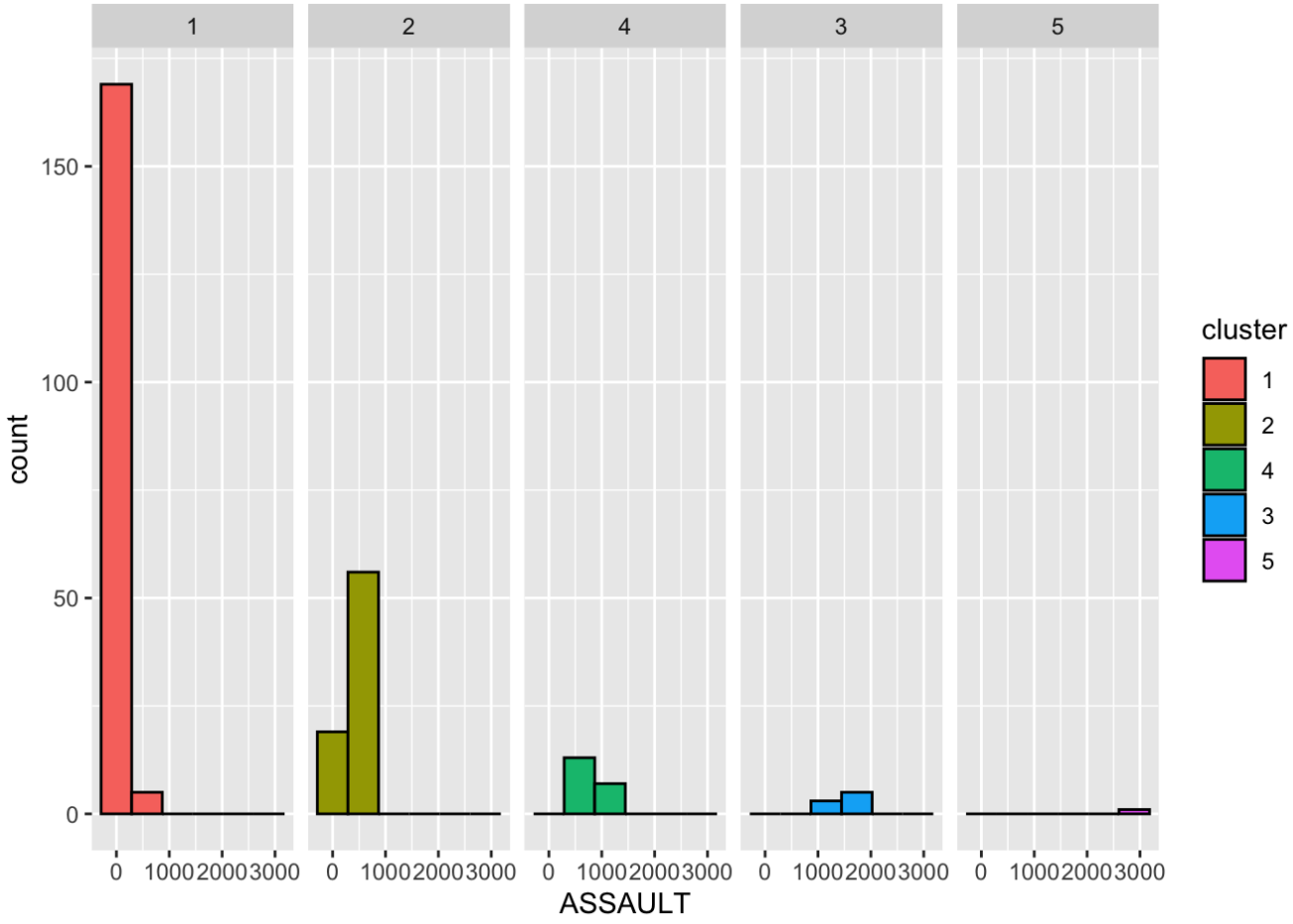



```
ggplot(data = to_clust, aes(x = "", y = ROBBERY)) + geom_boxplot() + facet_grid(~c
luster)
```



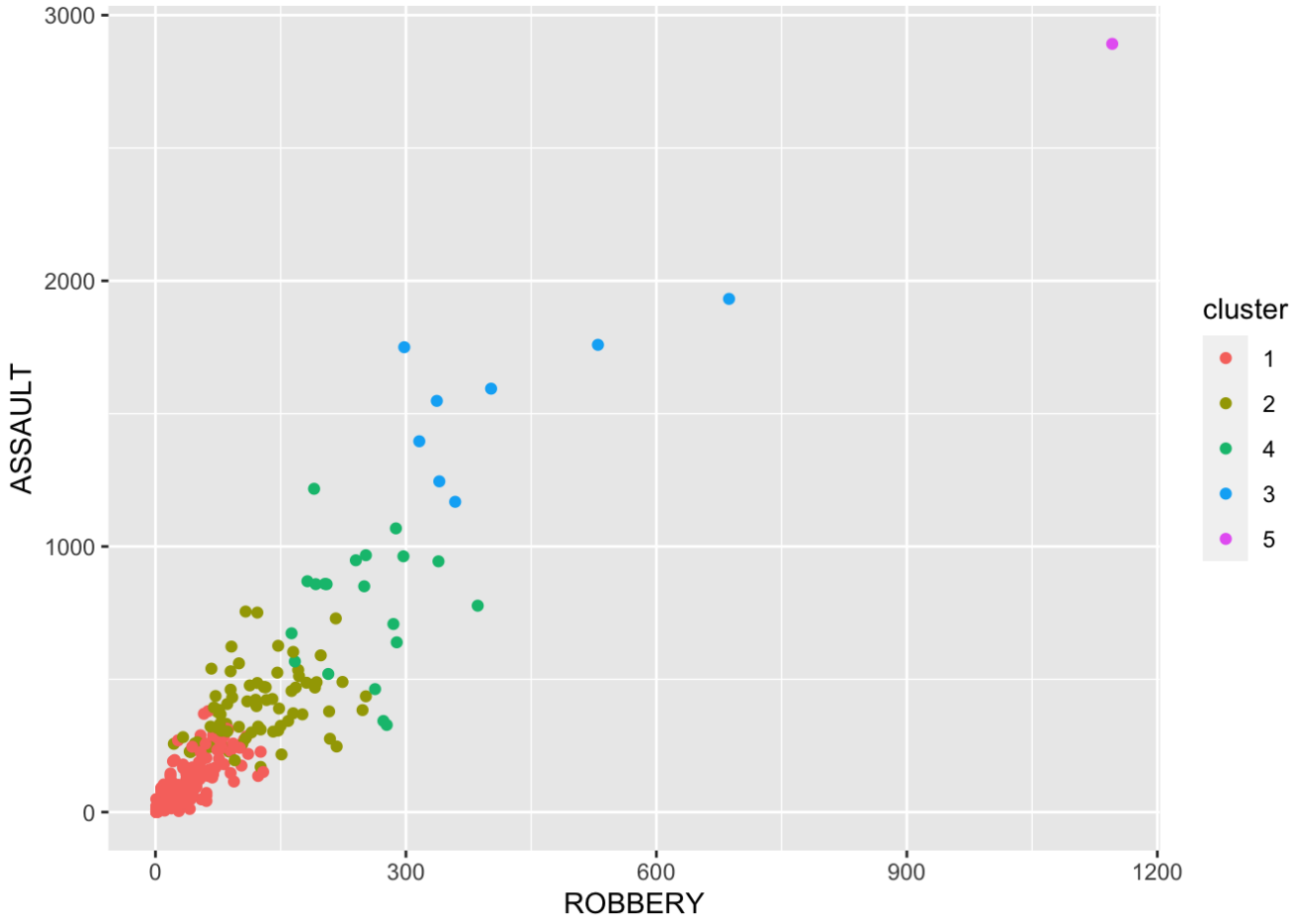

Как мы видим, распредления перстпулений по разным кластерам сильно отличаются как и по медианному значению, так и по разбросу данных. То же явление мы уже наблюдали выше.

```
ggplot(data = to_clust, aes(x = ASSAULT , fill = cluster)) + geom_histogram(bins = 6, col = "black") + facet_grid(~cluster)
```




Построим еще и гистограмму по данным по нападениям. Ситуация такая же, как мы и описали раннее: каждый кластер отражает свой уровень преступности в последовательности 1 - 2 - 4 - 3 - 5, где 1 кластер - самый безопасный, а 5 - самый опасный. Отличия в распределении преступлений по кластерам сильно заметны.

```
ggplot(data = to_clust, aes(x = ROBBERY, y = ASSAULT)) + geom_point(aes(color = cl
uster))
```




Еще один наглядный график. Мы построили диаграмму рассеивания, где по оси x - значения грабежей, а по оси y - нападений. Цвет каждой точки отражает ее кластер. Заивисимость между грабежами и нападениями сильно видна, а распределение точек по цветам выглядит достаточно логично и адекватно: точек, заходящие в районы точек с другим цветом, совсем мало.

# Формальные тесты

Воспользуемся формальным критерием Краскела-Уоллиса, нулевая гипотеза которго состоит в том, что данные в выборках взяты из одного распределения (медианы распределений равны):

```
kruskal.test(to_clust$ASSAULT ~ to_clust$cluster)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$ASSAULT by to_clust$cluster
## Kruskal-Wallis chi-squared = 180.57, df = 4, p-value < 2.2e-16
```

```
kruskal.test(to_clust$BURGLARY ~ to_clust$cluster)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$BURGLARY by to_clust$cluster
## Kruskal-Wallis chi-squared = 140.86, df = 4, p-value < 2.2e-16
```

```
kruskal.test(to_clust$HOMICIDE ~ to_clust$cluster)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$HOMICIDE by to_clust$cluster
## Kruskal-Wallis chi-squared = 120.32, df = 4, p-value < 2.2e-16
```

```
kruskal.test(to_clust$LARCENY ~ to_clust$cluster)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$LARCENY by to_clust$cluster
## Kruskal-Wallis chi-squared = 150.95, df = 4, p-value < 2.2e-16
```

```
kruskal.test(to_clust$RAPE ~ to_clust$cluster)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$RAPE by to_clust$cluster
## Kruskal-Wallis chi-squared = 142.44, df = 4, p-value < 2.2e-16
```

```
kruskal.test(to_clust$ROBBERY ~ to_clust$cluster)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$ROBBERY by to_clust$cluster
## Kruskal-Wallis chi-squared = 163.16, df = 4, p-value < 2.2e-16
```

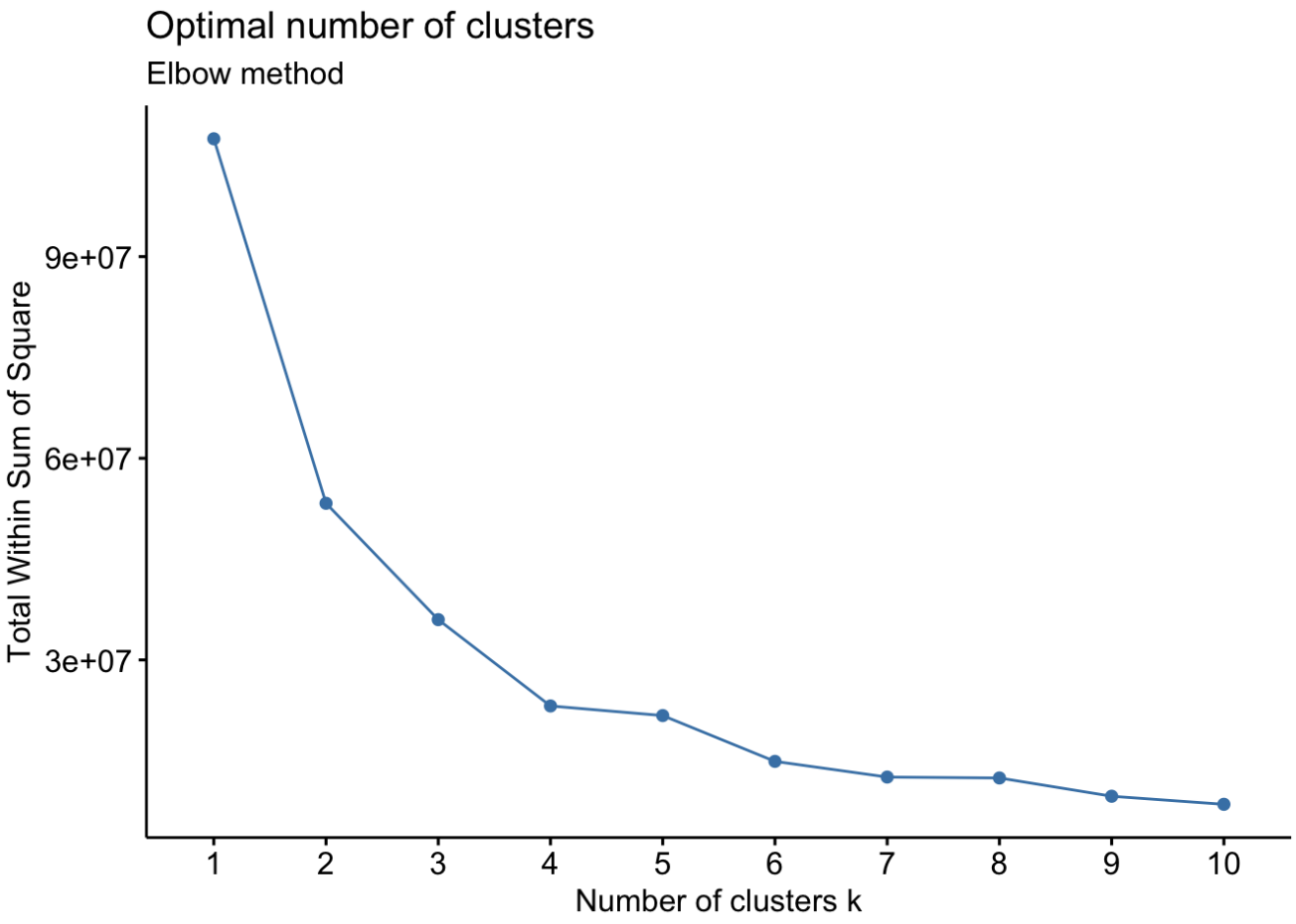
p-value во всех тестах очень мало, поэтому у нас есть основания на любых адекватных уровнях значимости отвергнуть нулевую гипотезу. Медианы показателей по кластерам не равны.

# Уточнение числа кластеров

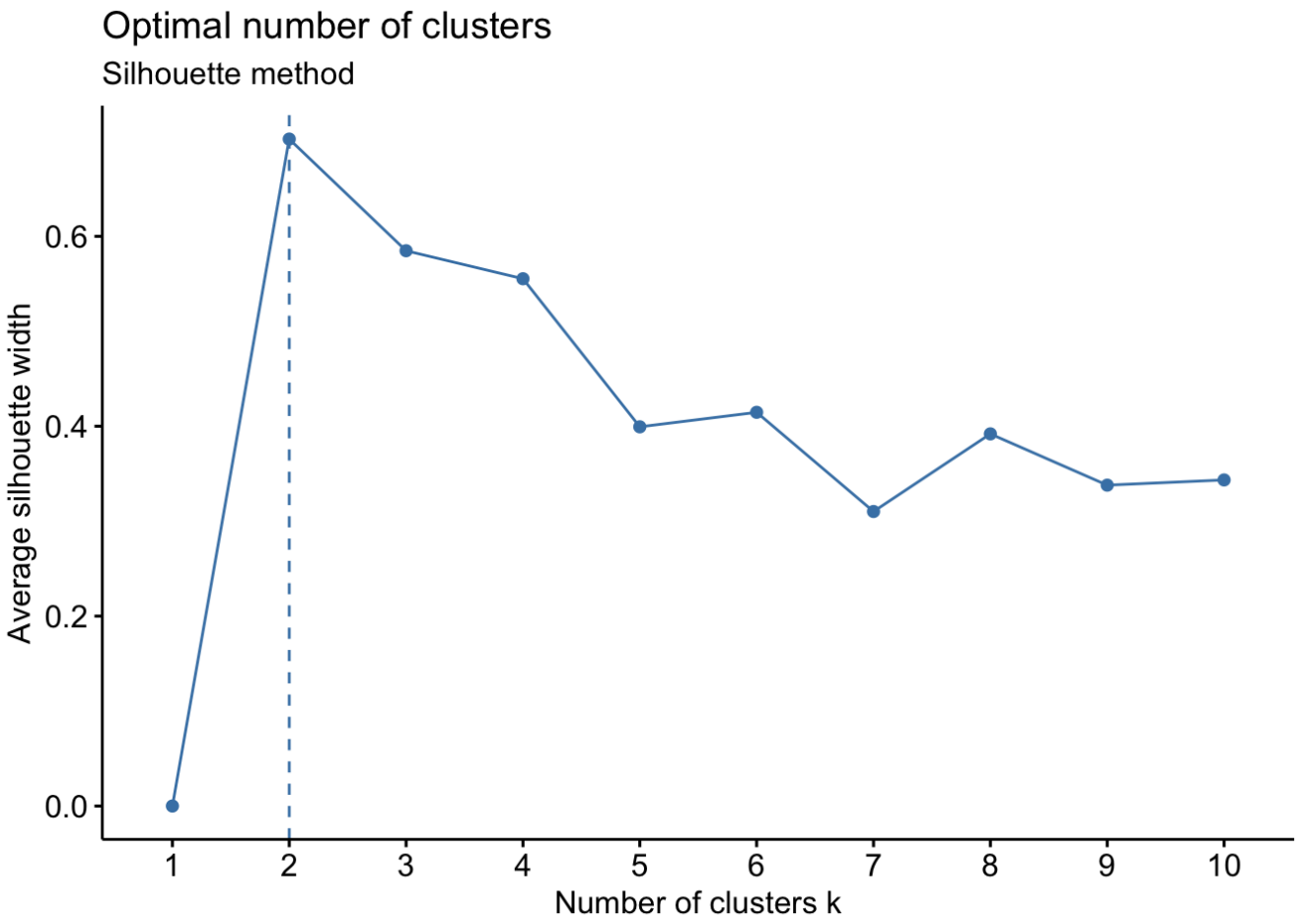
```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.g
l/ve3WBa
```

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




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



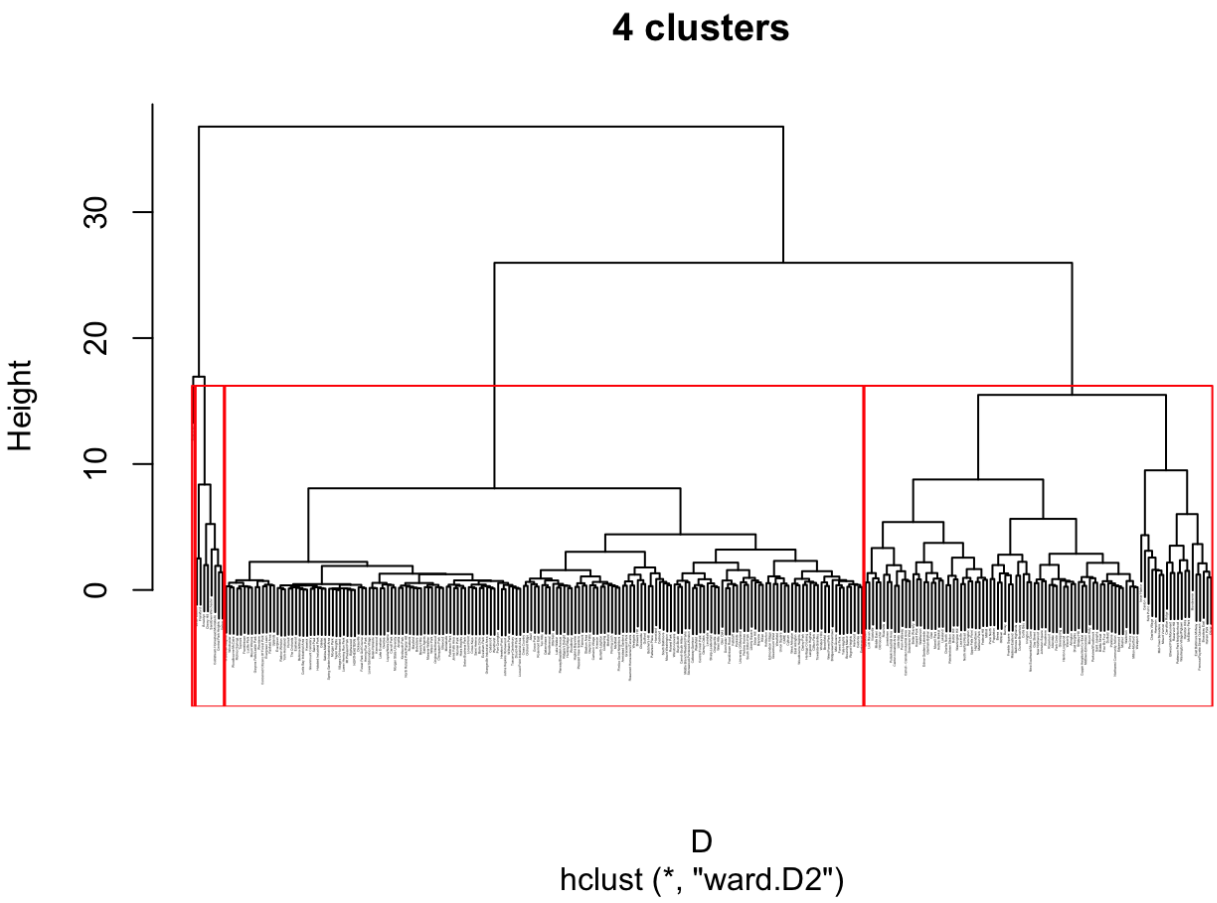

Посмотрим на первый график. “Колено” сгибается на значении 4 кластеров, врпочем, при 2-ч оно также достаточно сильно сгибается. Вариант с двумя кластерами нам не пойдет: они получатся слишком общие, а вот 4 кластера, судя по всему, будут лучше, чем 5.

Посмотрим на второй график. Силуэтный метод выдает нам 2 как оптимальное число кластеров. Нам этого все еще маловато.

# Дополнительная часть: так 4 или 5?

Мне было турдно решить, что же все-таки лучше брать: 4 или 5 кластеров. Кажется, что в нашем случае с существованием кластера, состоящего из 1 района с аномально высокими значениями по всем переменным, будет логично разделить остальные 277 районов на 4, а не на 3 кластера. Мне кажется, что разделение оставшихся районов на 3 кластера будет слишком общим. Но чтобы не опираться только на собственные содержательные доводы, предлагаю так же методом Варда разделить исходный массив на 4 кластера, вывести по ним некоторые из описательных статистик и подумать, стоит ли оставлять 5 кластеров или выбрать итоговое значение, равное 4.

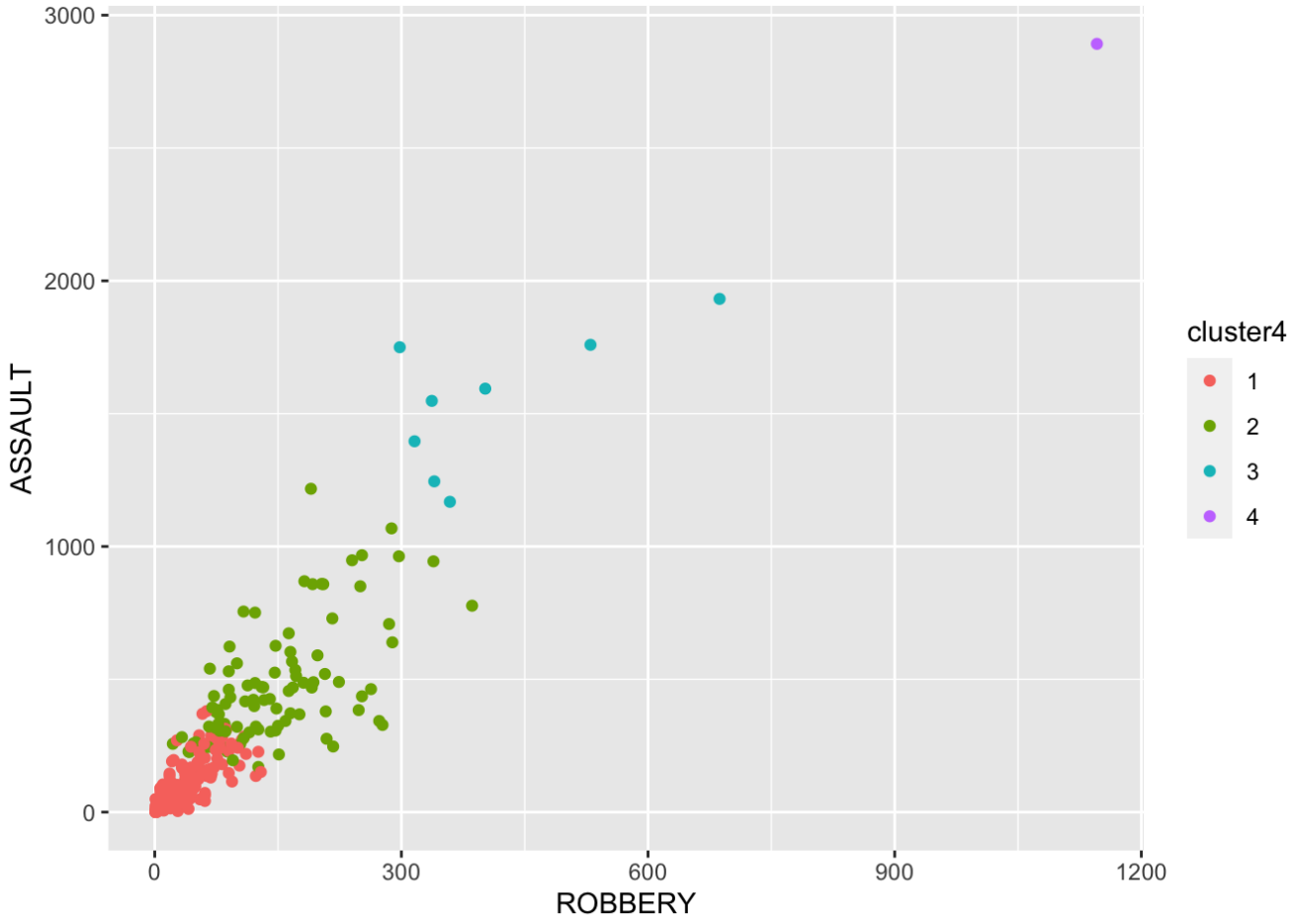
```
cluster4 <- cutree(hc, k = 4)
to_clust$cluster4 <- factor(cluster4)

plot(hc, cex = 0.1, main = "4 clusters")
rect.hclust(hc, k = 4, border="red")
```




```
to_clust %>% group_by(cluster4) %>% summarise_at(vars(ASSAULT:ROBBERY), mean)
```

```
## # A tibble: 4 × 7
##   cluster4 ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>      <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1           111.     80.8     1.79    161.  2.70    38.5
## 2 2           468.    255.      9.12    616.  9.54   147.
## 3 3          1549     759      34.2    1194. 31.1    409.
## 4 4          2892     486      19      4764  83     1146
```

```
ggplot(data = to_clust, aes(x = ROBBERY, y = ASSAULT)) + geom_point(aes(color = cl
uster4))
```

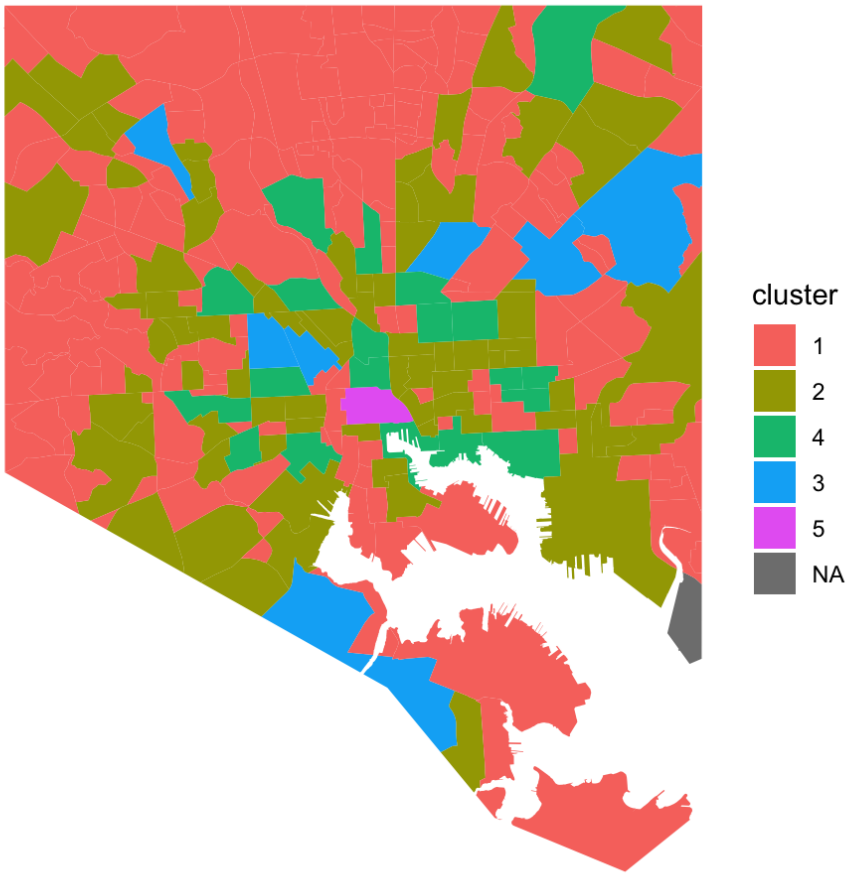



Средние значения в 3 первых кластерах различаются прмерно в 4 раза. Диаграмма рассеиваний для значений нападения-грабежи тоже вышла достаточно логичной и ровной. Важно только решить, нравится ли нам отличие в примерно 4 раза между показателями.

Хоть по формальным тестам нам выдало оптимальное значений кластеров равное 4 (2 нам точно не подходит, слишком общее значений), **я бы предпочла остановиться на 5**. Карта города по преступности чаще всего строится для оценки, например, стоимости недвижимости в том или ином районе, или используется как фактор при принятии решения о переселении в тот или иной район. В таком случае, кажется, что разделить карту на 3 кластера (все еще пока закрывает глаза на криминальный центр), где средние показатели по переменным отличаются в 4 раза - достаточно общий выход из проблемы деления районов на кластеры. 5 кластеров, хоть они уже более детализированные, но все-таки отображающие больше важной информации, особенно когда речь идет о не просто индикаторе уровня инфраструктуры и количества хипстерских кофеен, продающих матчу, а об уровне преступности. Избегать детальности в серьезном вопросе - достаточно пренебрежительно.

# Кластеры и география

```
crime$cluster <- to_clust$cluster
full <- fortified %>% left_join(. , crime, by=c("id"="Neighborhood"))
ggplot() + geom_polygon(data = full,
           aes(fill = cluster, x = long, y = lat, group = group)) +
           theme_void() + coord_map()
```




Самый опасный район в центре - об этом мы уже сказали и показали, почему аномальная преступность сосредоточена в центре города.

Самые безопасные районы - красные - расположены в основном на окраинах города. Скорее всего, это новые жилые кварталы, поэтому уровень преступности в них минимальный. На севере Балтимора, где, кажется, сосредоточено наибольшее число безопасных районов, находятся самые известные исторические районы.

Юг Балтимора - смешанные жилые и промышленные районы. Уровень преступности там разный: есть и красные (безопасные) районы, и синие - опасные, и пару грязно-зеленых - с допустимой преступностью.

Северо-восточный и восточные районы жилые. Большой синий район, помеченный как “опасный” - это район Франкфорд, который является одним из самых густонаселенных районов. Распределение районов там по нашей карте тоже неоднородное, схоже с распределением на Юге.

Западные районы Балтимора кажутся также наименее опасными: много красных районов, чуть меньше грязно-зеленых, они тоже преимущественно жилые.

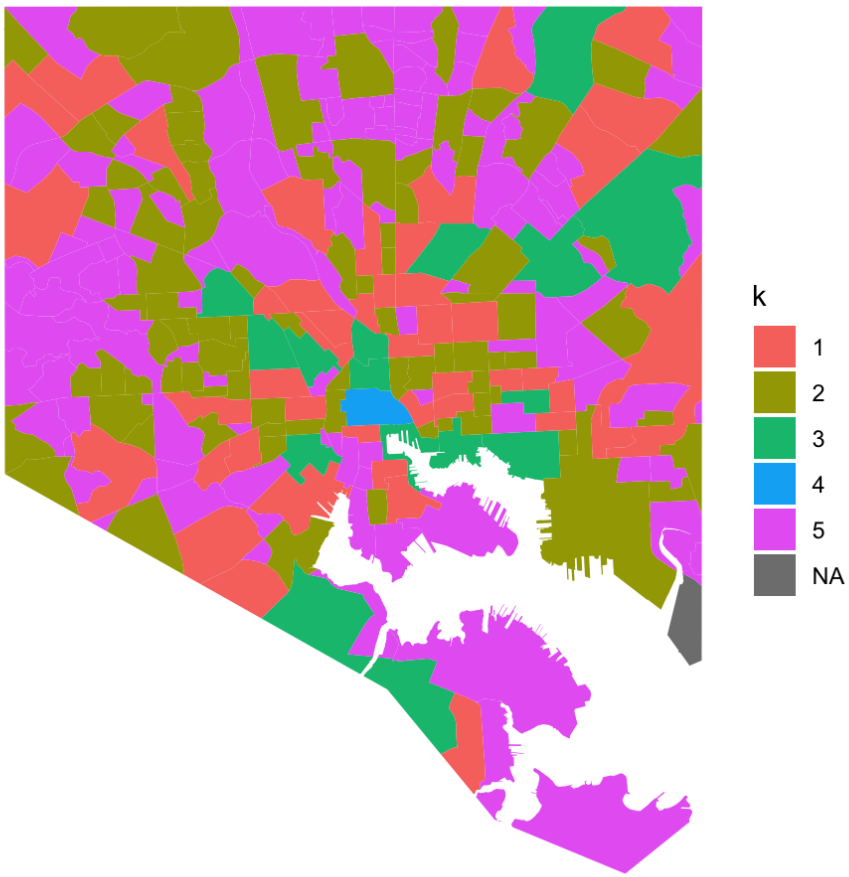
Откровенно говоря, заметить какую-то зависимость между расположением района и уровнем преступности в нем не так просто. Кажется, что в разных жилых районах распредление примерно похожее: где-то чуть больше перступности, где-то чуть меньше. Районны, классифицируемые как средне опасные (зеленые) и сильно опасные (синие) расположены тоже достаточно случайно на первый взгляд. Для того, чтобы сделать более обоснованные выводу по распределению преступности в районах от их расположения нужно больше знать о районах Балтимора: какие районы новые, какие старые, какие районы более и менее этнически разнообразны, какие районы - центры молодежной жизни, а какие районы утопают в бизнес-центрах.

# K-means

```
kclust <- kmeans(to_clust[1:6], 5)
to_clust$k <- factor(kclust$cluster)
```

```
cluster01_k <- to_clust %>% filter(k == 1)
cluster02_k <- to_clust %>% filter(k == 2)
cluster03_k <- to_clust %>% filter(k == 3)
cluster04_k <- to_clust %>% filter(k == 4)
cluster05_k <- to_clust %>% filter(k == 5)
```

```
crime$k <- to_clust$k
full <- fortified %>% left_join(. , crime, by=c("id"="Neighborhood"))
ggplot() + geom_polygon(data = full,
           aes(fill = k, x = long, y = lat, group = group)) +
           theme_void() + coord_map()
```




Деление на кластеры с помощью k-means сильно отличается.

Мы получили очень неоднородную карту, поэтому выделить завимость преступности от расположения района будет снова очень сложно выделить.

# Интерпретация кластеров, полученных с помощью k-means

```
to_clust %>% group_by(k) %>% summarise_at(vars(ASSAULT:ROBBERY), mean)
```

```
## # A tibble: 5 × 7
##   k     ASSAULT BURGLARY HOMICIDE LARCENY  RAPE ROBBERY
##   <fct>   <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl>
## 1 1       558.     295.      9.52    702. 11.0    176.
## 2 2       277.     167.      6.14    304.  6.07    84.0
## 3 3      1106      601.     19.4    1499. 22.7    344.
## 4 4      2892      486      19      4764  83     1146
## 5 5        80.0     58.1     1.55    119.  2.03    27.3
```

# 1 кластер

```
cluster01_k[1:5, 1:6]
```

```
##                     ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Baltimore Highlands     729      199        7     677   17     216
## Barclay                 512      260       17     543   10     172
## Better Waverly          535      353       11     839    5     171
## Bolton Hill             247      216        1     919    6     217
## Broadway East          1217      475       28     620   16     190
```

В первом кластере находятся районы с *нормальным уровнем преступности*. Это кластер не с райоными, содержащими абсолютный минимум по числу преступлений, а кластер с незначительным, допустимым повышением этого уровня.

Среднее число нападений - 277, краж со взломом - 167, убийств - 6, хищений - 304, изнасилований - 6, грабежей - 84.

Судя по карте, районы с таким уровнем преступности расположены неоднородны и независимо, распределены по всей карте города, что нормально для районов с нормальным уровнем преступнсоти :)

# 2 кластер

```
cluster02_k[1:5, 1:6]
```

```
##                   ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Abell                 147      122        1     269    1      90
## Allendale             368      255       13     283    6      78
## Arlington             417      180       11     301    7     110
## Armistead Gardens     289      192        1     305    5      54
## Ashburton             115      249        6     230    1      94
```

*Самый безопасный район*. Достигает абсолютного минимума по всем показателям преступлений.

Среднее число нападений - 80, краж со взломом - 58, убийств - 1.5, хищений - 118.5, изнасилований - 2, грабежей - 27.

Судя по карте, эти районы снова раскинуты по всему городу с одним замечанием: их практически нет в центре, а те, которые есть - очень малы по размеру. Это понятно: районы с совсем малым

количеством преступлений вряд ли будут находиться в центре, скрее, это будут районы с новыми застройками или те, где хорошо работает отделение полиции.

# 3 кластер

```
cluster03_k[1:5, 1:6]
```

```
##                                ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Belair-Edison                     1759     1129       40    1919   31      530
## Brooklyn                          1594      852       30    1061   45      402
## Canton                             639      637        2    2174   10      289
## Cherry Hill                       1548      806       20    1071   34      337
## Coldstream Homestead Montebello   1245      592       43     833   17      340
```

В нашем рейтинге районов районы из 3 кластера относятся к категории *сильно преступных* районов. Фактически, это самые опасные районы города (не считая тот-самый-аномальный район в виде центра города).

Среднее число нападений - 1106, краж со взломом - 601, убийств - 19, хищений - 4764, изнасилований - 23, грабежей - 343.5.

В их расположении уже более видна неоднородность. На западе Балтимора совсем нетс ильно опасных районов, немного в центре, немного на Юге (хотя на Юге таких районов 2, они занимают наибольшую площадь), немного на северо-востоке.

# 4 кластер

```
cluster04_k[, 1:6]
```

```
##          ASSAULT BURGLARY HOMICIDE LARCENY RAPE ROBBERY
## Downtown    2892      486       19    4764   83    1146
```

В 4 кластер входит 1 район - и это центр города. Это единственный район с огромным количеством всех видов преступлений, средние значения в таблице. Центру города достается абсолютная победа в номинации *самый опасный район Балтимора*.

! С помощью k-means немного подправилась ситуация с количество убийств в районах из кластера с опасными районами иц ентром. Напомним, что при кластеризации методом Варда количество убийств стало единственным показателем, по которому кластер самых опасных районов обогнал центр. В этот раз их значения почти одинаковы - 19.

# 5 кластер

В 5 кластере собрались *опасные* районы. Они ниже по значениям, чем самые опасные районы, но уже в среднем в 2 раза превышают значения спокойных районов.

Среднее число нападений - 558, краж со взломом - 295, убийств - 9.5, хищений - 701.5, изнасилований - 11, грабежей - 176.

На карте они расположены достаточно равномерно по всему Балтимору.