# K-means

## Теория и практика применения

Promedia consulting

Київ, 2017

© Voronoi Shelf by Marc Newson

# **Кирилл** Захаров

Директор аналитического департамента

**8+**
лет опыта работы с данными

**20**
разработанных приложений

**100**
социологических исследований

**4**
языка программирования

**200**
тренингов по разным вопросам

Promedia consulting

# Что будем **обсуждать?**

Назначение кластерного анализа

История метода К-средних

Как работает метод К-средних

Особенности различных алгоритмов

Преимущества и недостатки метода

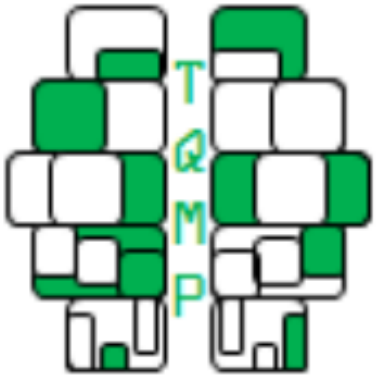
Основные этапы работы с методом К-средних

Особенности статистического вывода при применении метода

Promedia consulting

# Где **почитать?**

**Zakharov, K.** (2016).
Application of k-means clustering in psychological studies.
The Quantative Methods for Psychology, 12(2), 87-100,
**doi:** 10.20982/tqmp.12.2.p087

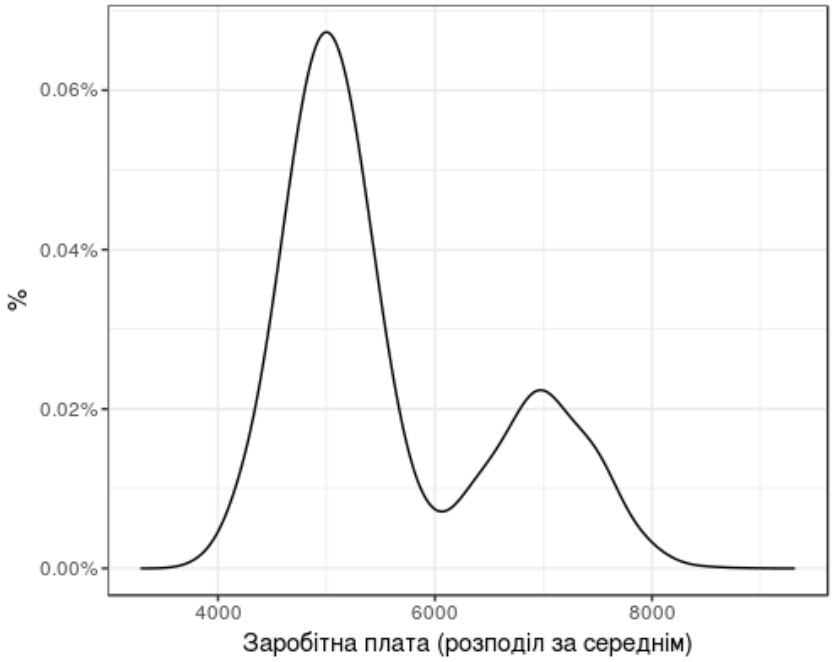



Promedia consulting

# **Назначение** кластерного анализа

Представим, что у нас есть группа респондентов с известным доходом. Наша задача выделить **социальные категории** для дальнейшего анализа

```
library(ggplot2)
library(scales)

set.seed(123)

data <- round(c(rnorm(5000, 5000, 400),
                rnorm(2000, 7000, 500)), 0)

dd <- with(density(data), data.frame(x, y))

qplot(x, y, data = dd, geom = "line")+
 theme_bw() +
 xlab("Зарплата (распределение по середнему)") +
 ylab("%") + scale_y_continuous(labels = percent)
```

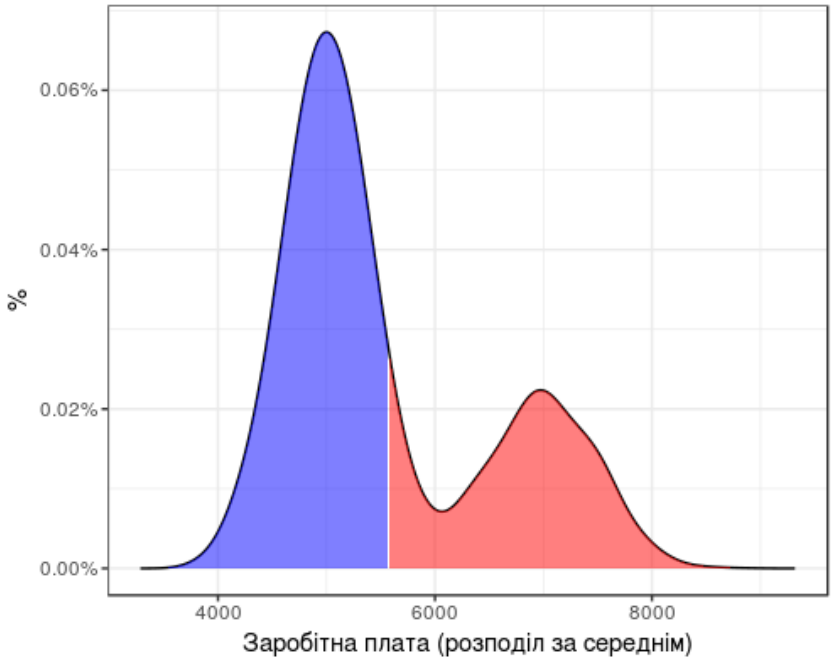



Promedia consulting

# **Назначение** кластерного анализа

Попробуем выделить категории по **среднему значению**

```
qplot(x, y, data = dd, geom = "line") +

  geom_ribbon(data = subset(dd, x > mean(data)),
    aes(ymax = y), ymin = 0, fill = "red", alpha = 0.5) +

  geom_ribbon(data = subset(dd, x < mean(data)),
    aes(ymax = y), ymin = 0, fill = "blue", alpha=0.5) +

  theme_bw() + xlab("Заробітна плата (розподіл за
    середнім)") + ylab("%") +
  scale_y_continuous(labels = percent)
```

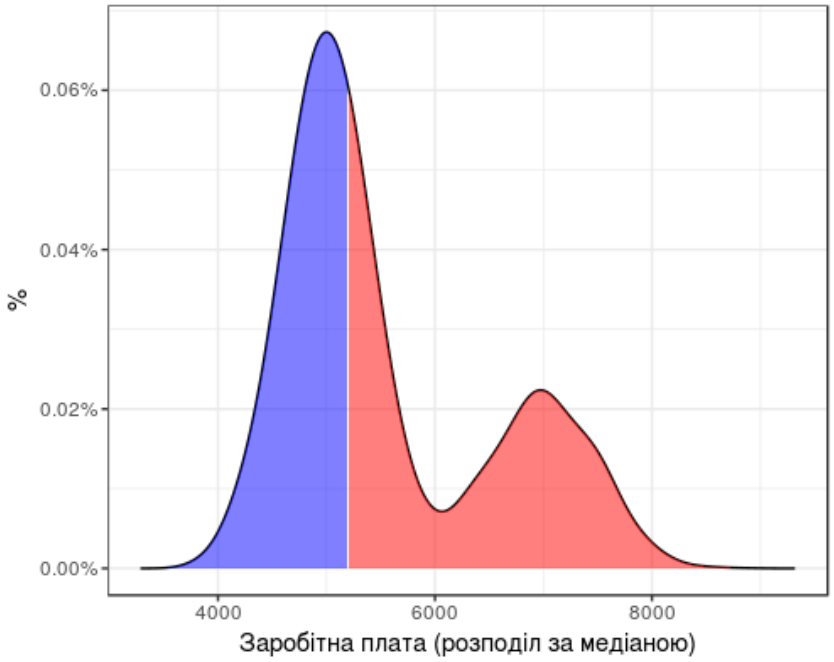



Promedia consulting

# **Назначение** кластерного анализа

Попробуем выделить категории по **медиане**

```
qplot(x, y, data = dd, geom = "line") +

  geom_ribbon(data = subset(dd, x > median(data)),
    aes(ymax = y), ymin = 0, fill = "red", alpha = 0.5) +

  geom_ribbon(data = subset(dd, x < median(data)),
    aes(ymax = y), ymin = 0, fill = "blue", alpha=0.5) +

  theme_bw() + xlab("Заробітна плата (розподіл за
    медіаною)") + ylab("%") +
  scale_y_continuous(labels = percent)
```

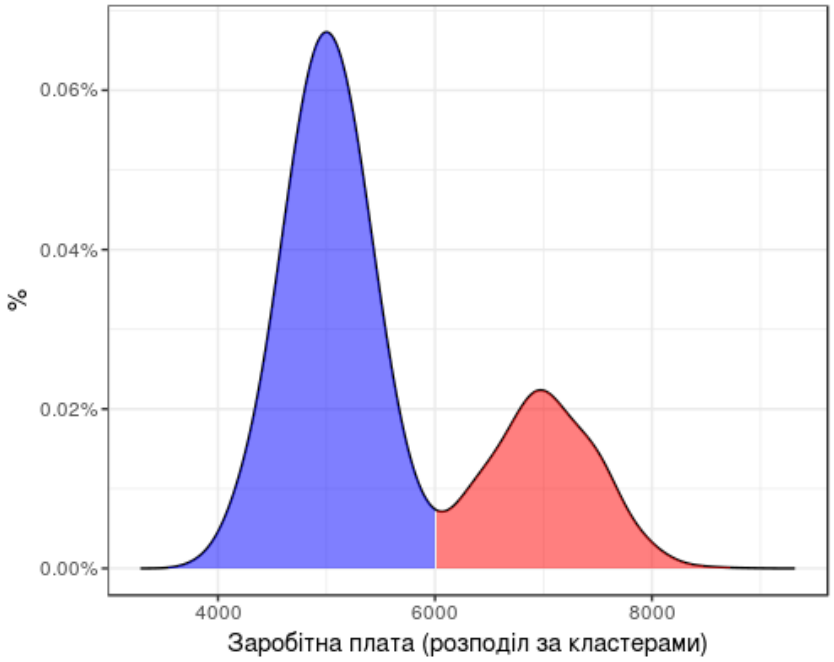



Promedia consulting

# **Назначение** кластерного анализа

Попробуем выделить категории **методом К-средних**

```
set.seed(123)
cluster = kmeans(data, centers = 2)

qplot(x, y, data = dd, geom = "line") +

  geom_ribbon(data = subset(dd,
     x > min(data[cluster$cluster == 2]))),
     aes(ymax = y), ymin = 0, fill = "red", alpha = 0.5) +
  geom_ribbon(data = subset(dd, x < median(data)),
     aes(ymax = y), ymin = 0, fill = "blue", alpha=0.5) +
  theme_bw() + xlab("Заробітна плата (розподіл за
     кластерами)") + ylab("%") +
  scale_y_continuous(labels = percent)
```




Promedia consulting

# **Назначение** кластерного анализа

Что если нам необходимо разделить выборку по **нескольким переменным** и мы используем градации переменных?

Количество градаций: **2 × 2 × 3 × 3 × 4 × 8**

Количество групп: **2 4 12 36 144 1152**

Численность выборки для статистичесского вывода:

**5760?**

Promedia consulting

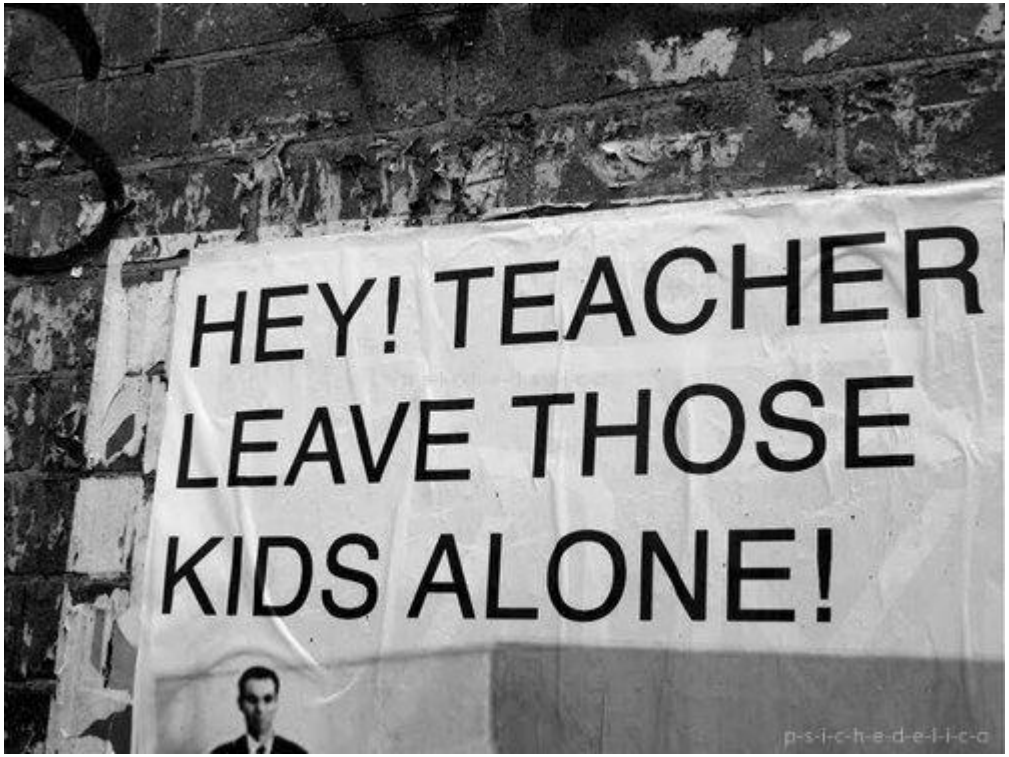
# **Назначение** кластерного анализа

Метод K-средних позволяет строить так называемые **эмпирические классификации**, в основе которых лежит количественная обработка опытных данных.

**Цель кластерного анализа** заключается в нахождении существующих в данных структур — кластеров. При этом объекты в каждом кластере должны быть максимально схожи между собой и отличаться от объектов в других кластерах.

Promedia consulting

# **Назначение** кластерного анализа




Метод K-средних является одним из методов классификации **без учителя**

Promedia consulting

# **Назначение** кластерного анализа

1. Для **эксплораторного анализа** и построения классификаций в исследовательской практике и интеллектуальном анализе данных.
2. Для редукции (**уменьшения сложности**) данных.
3. Как **начальный шаг** для более сложных в вычислительном плане алгоритмов, который дает приблизительное разделение данных как новые начальные точки (уменьшение зашумления в наборе данных).

Л. Мориссетт и С. Картье

# **История** метода K-средних

**1939**

**Robert Tryon**
Cluster analysis

**1965**

**Edward Forgy**
Cluster analysis of multivariate data: efficiency versus interpretability of classifications

**1967**

**James MacQueen**
Some methods for classification and analysis of multivariate observations

**1982**

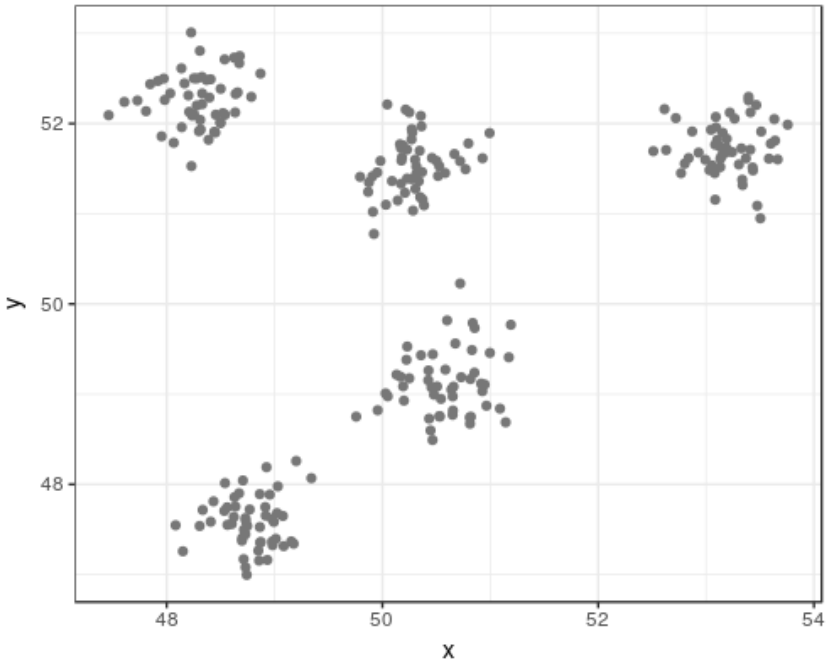
**Stuart Lloyd**
Least square quantization in PCM

Promedia consulting

# **Как работает** K-средних

Представим, что у нас есть ряд объектов многомерном пространстве

```
set.seed(1)
k = 5; n = 50
data <- data.frame(
    x = rnorm(k, 50 ,2), y = rnorm(k, 50, 3))

for (row in 1:nrow(data)) {
  data <- rbind(data, data.frame(
     x = rnorm(n, data[row, ]$x, 0.3),
     y = rnorm(n, data[row, ]$y, 0.3)))
}

ggplot(data, aes(x = x, y = y)) +
  geom_point(color = '#777777') +
  theme_bw()
```

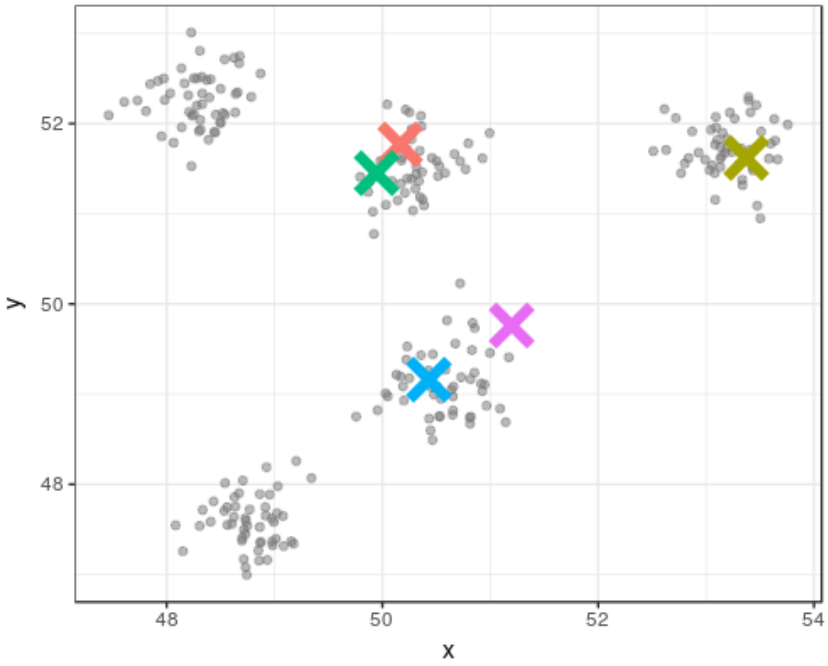



Promedia consulting

# **Как работает** K-средних

В общем виде **смысл алгоритма** метода K-средних заключается в реализации следующих шагов:

1. Случайным образом выбираются K кластерных центров.

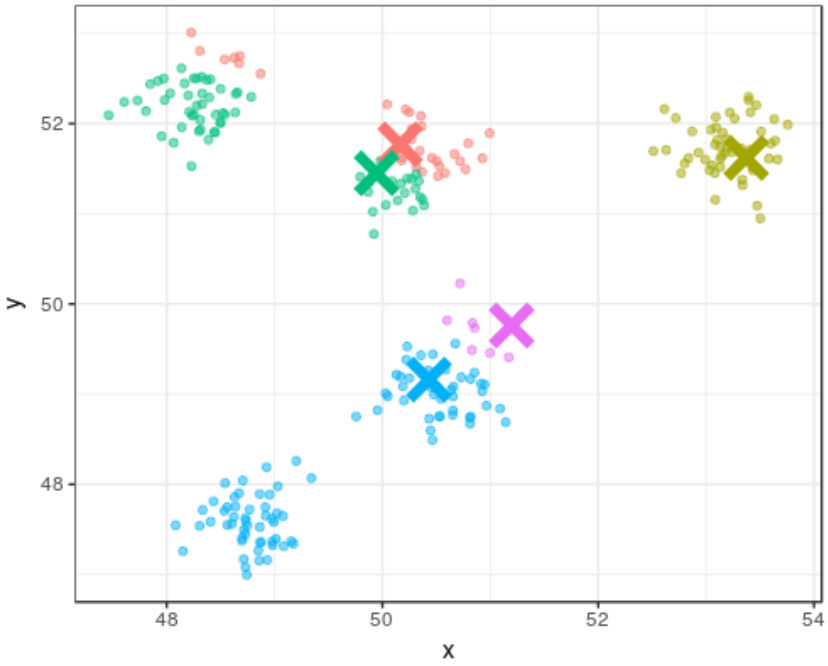



Promedia consulting

# **Как работает** K-средних

В общем виде **смысл алгоритма** метода K-средних заключается в реализации следующих шагов:

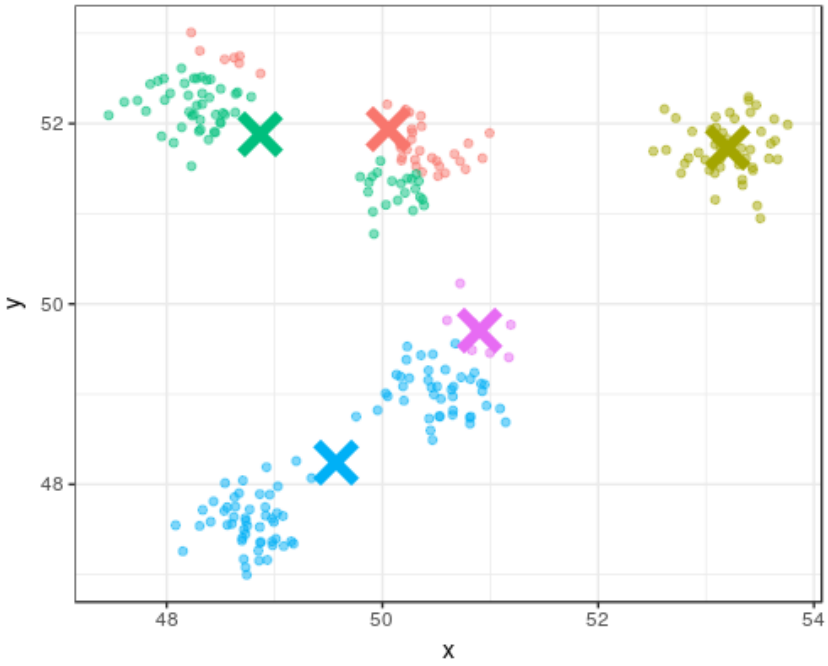
1. Случайным образом выбираются K кластерных центров.
2. Каждая точка назначается ближайшему центру кластеров.



# **Как работает** K-средних

В общем виде **смысл алгоритма** метода K-средних заключается в реализации следующих шагов:

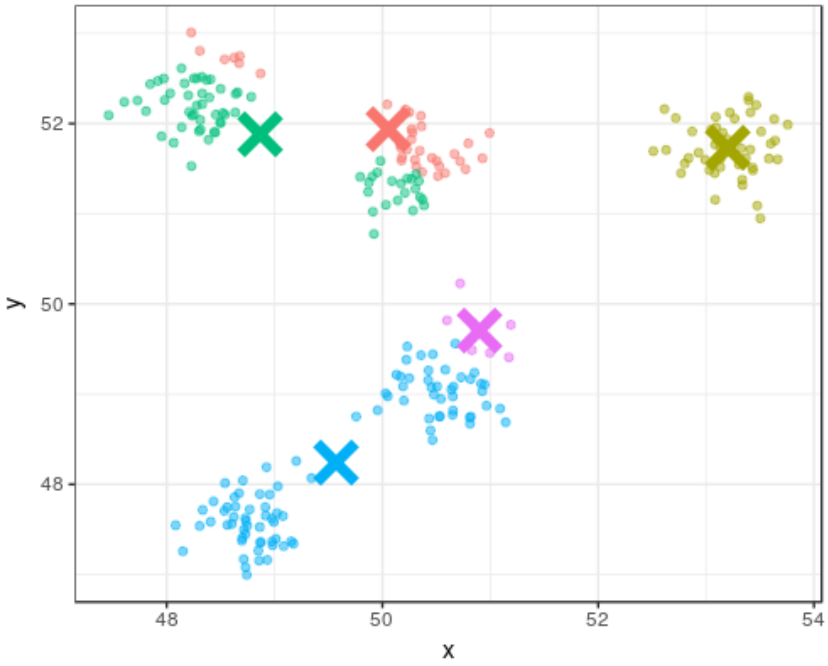
1. Случайным образом выбираются K кластерных центров.
2. Каждая точка назначается ближайшему центру кластеров.
3. Пересчитываются центры кластеров, используя текущее распределение кластеров.



# **Как работает** K-средних

В общем виде **смысл алгоритма** метода K-средних заключается в реализации следующих шагов:

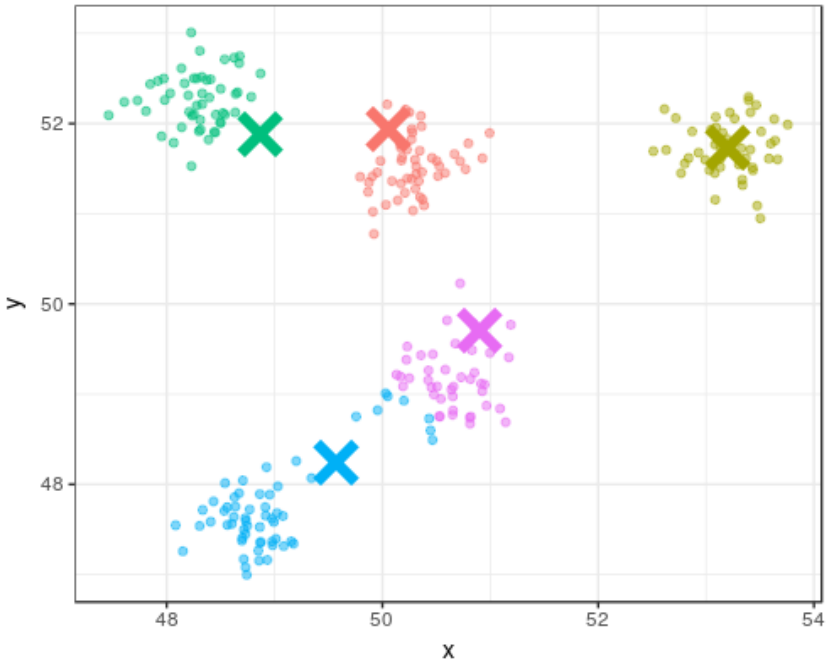
1. Случайным образом выбираются K кластерных центров.
2. Каждая точка назначается ближайшему центру кластеров.
3. Пересчитываются центры кластеров, используя текущее распределение кластеров.
4. Если критерий сходимости не удовлетворен, то идет возврат ко второму шагу.



# **Как работает** K-средних

В общем виде **смысл алгоритма** метода K-средних заключается в реализации следующих шагов:

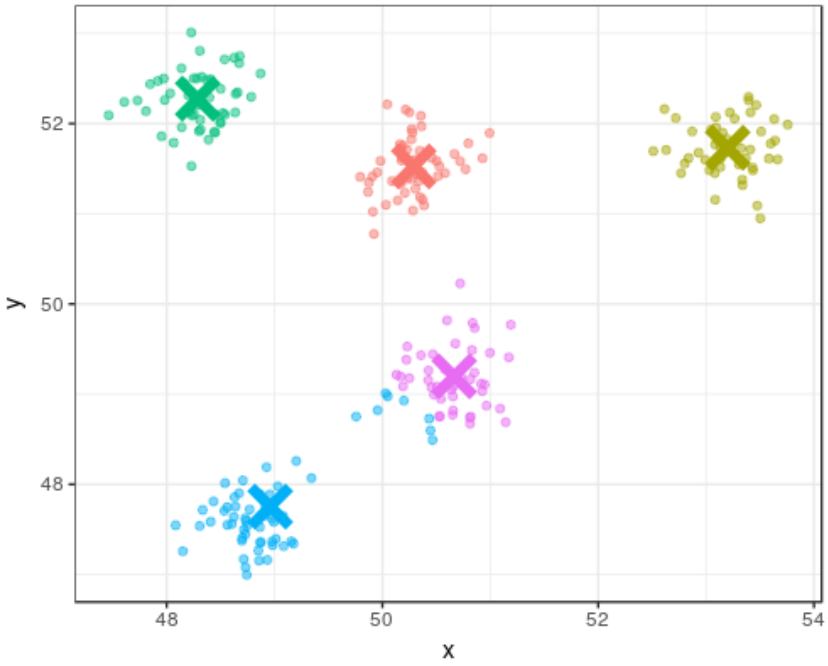
1. Случайным образом выбираются K кластерных центров.
2. Каждая точка назначается ближайшему центру кластеров.
3. Пересчитываются центры кластеров, используя текущее распределение кластеров.
4. Если критерий сходимости не удовлетворен, то идет возврат ко второму шагу.



# **Как работает** K-средних

В общем виде **смысл алгоритма** метода K-средних заключается в реализации следующих шагов:

1. Случайным образом выбираются K кластерных центров.
2. Каждая точка назначается ближайшему центру кластеров.
3. Пересчитываются центры кластеров, используя текущее распределение кластеров.
4. Если критерий сходимости не удовлетворен, то идет возврат ко второму шагу.

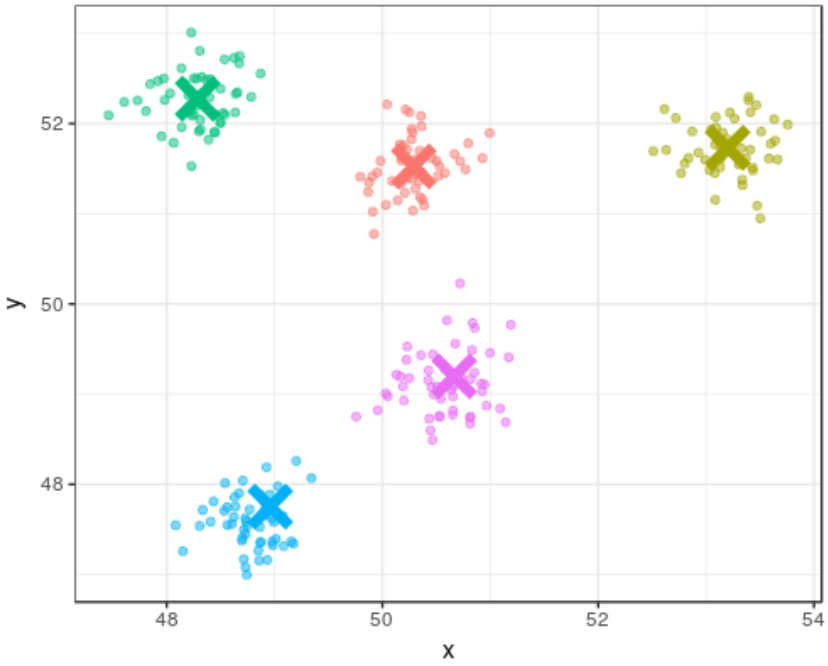



Promedia consulting

# **Как работает** K-средних

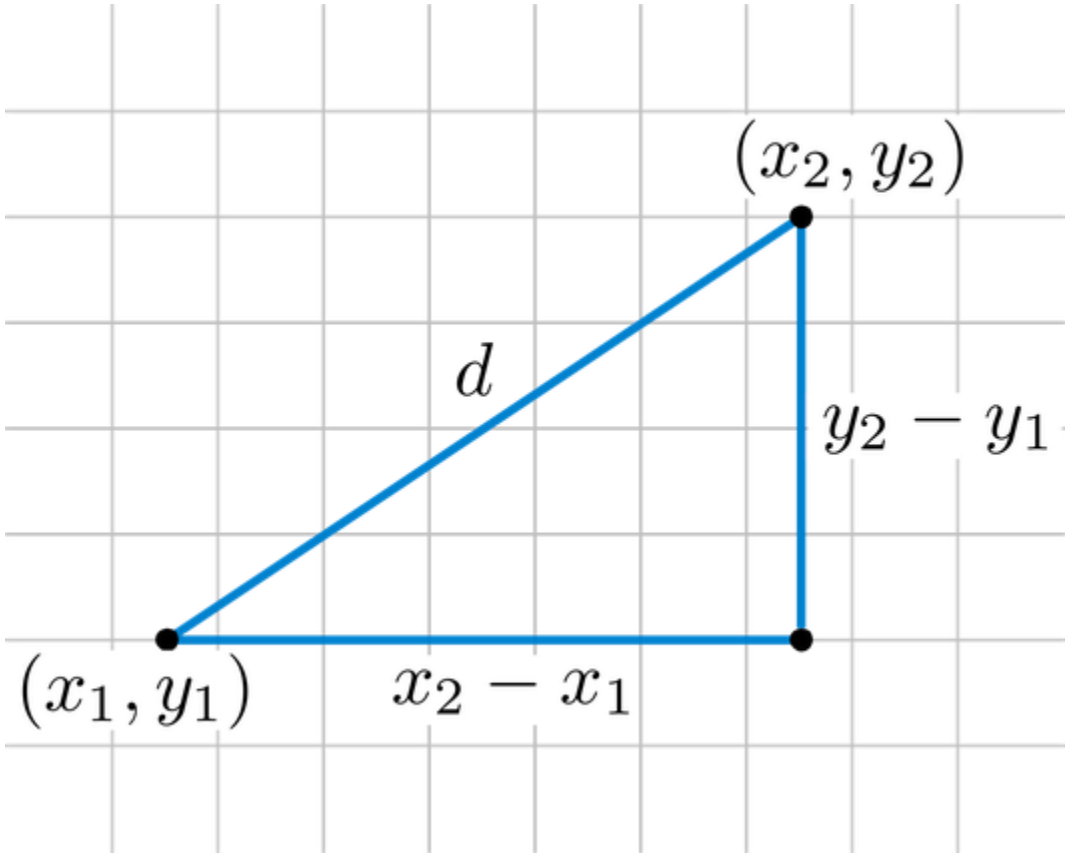
В общем виде **смысл алгоритма** метода K-средних заключается в реализации следующих шагов:

1. Случайным образом выбираются K кластерных центров.
2. Каждая точка назначается ближайшему центру кластеров.
3. Пересчитываются центры кластеров, используя текущее распределение кластеров.
4. Если критерий сходимости не удовлетворен, то идет возврат ко второму шагу.



# **Как работает** K-средних

Для расчетов расстояний между объектами чаще всего используется эвклидово расстояние



# **Специфика** алгоритмов

На практике используются различные разновидности алгоритмов:

| | |
|---|---|
| **Ллойда-Форджи** | Серийная модель центроид (использует все объекты одновременно). Начальные центроиды выбираются случайно их всех объектов. |
| **Случайного деления** | Исходные кластеры формируются случайным образом и из такой кластеризации рассчитываются начальные центроиды |
| **МакКуина** | Метод «бегущих средних». Центроиды пересчитываются после добавления каждого нового объекта. |
| **Хартигана-Вонга** | Критерий сходимости – минимальное значение внутригрупповой дисперсии кластеров. |

# **Премущества** K-средних

Простота алгоритма

Не требует вычисления и хранения матрицы расстояний

Возможность параллелизации

Линейная пространственная и временная сложность

# **Недостатки** K-средних

Метод всегда сходится, но может привести к нахождению локального минимума

В алгоритме Маккуина и Харигана и Вонга — решение чувствительно к порядку, в котором предъявляются точки, а при случайном выборе начальных точек могут создаваться пустые кластеры

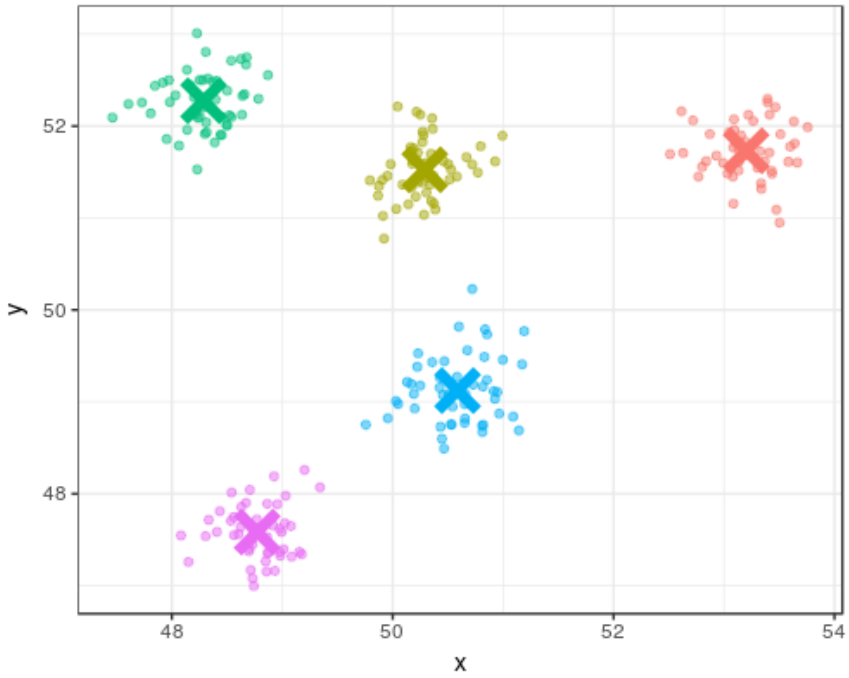
Выбор различных начальных центров приводит к различным решениям

Алгоритмы чувствительны к выбросам и зашумленным данным

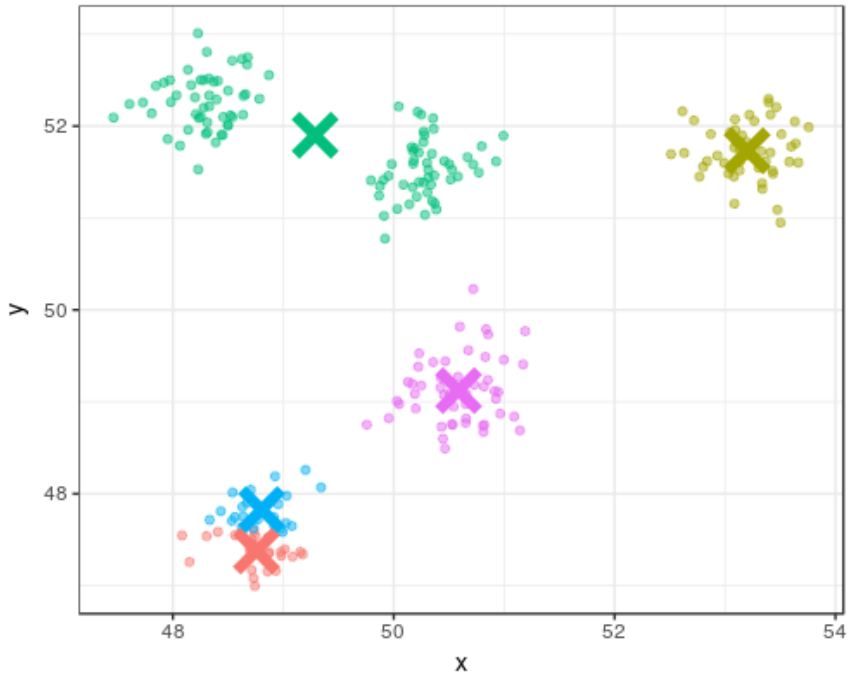
Метод стремится создать кластеры равного размера, даже если это неоптимально

# **Недостатки** K-средних

```
set.seed(1); clusters <- kmeans(data, centers = 5)
```

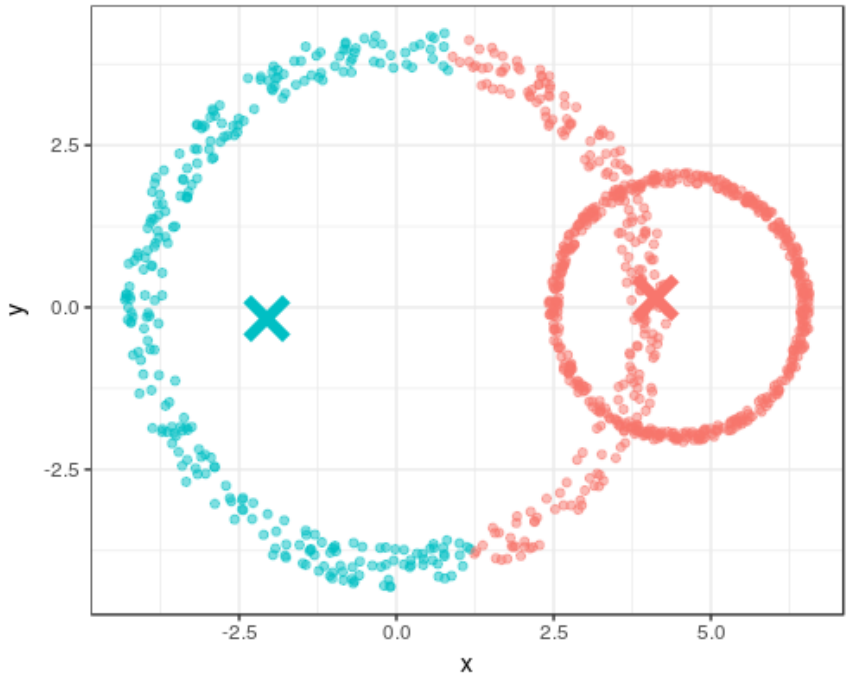


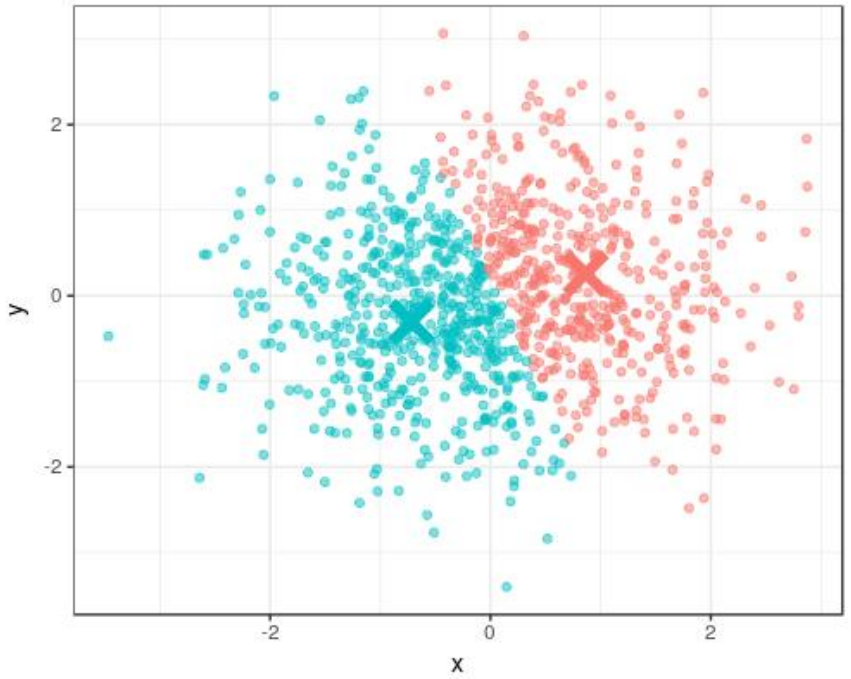

```
set.seed(2); clusters <- kmeans(data, centers = 5)
```




Разные начальные центры кластеров приводят к разным решениям. Решения могут быть неопитмальны.

# **Недостатки** K-средних



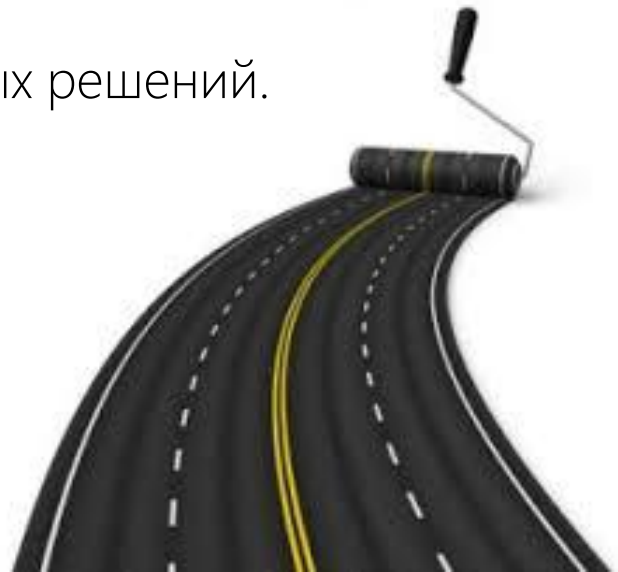

Равноразмерные кластеры неоптимальны.
Кластерный анализ выдаст результат даже в случае отстутсвия реальных кластеров.

# **Этапы применения** K-средних


Promedia consulting

# **Этапы применения** K-средних

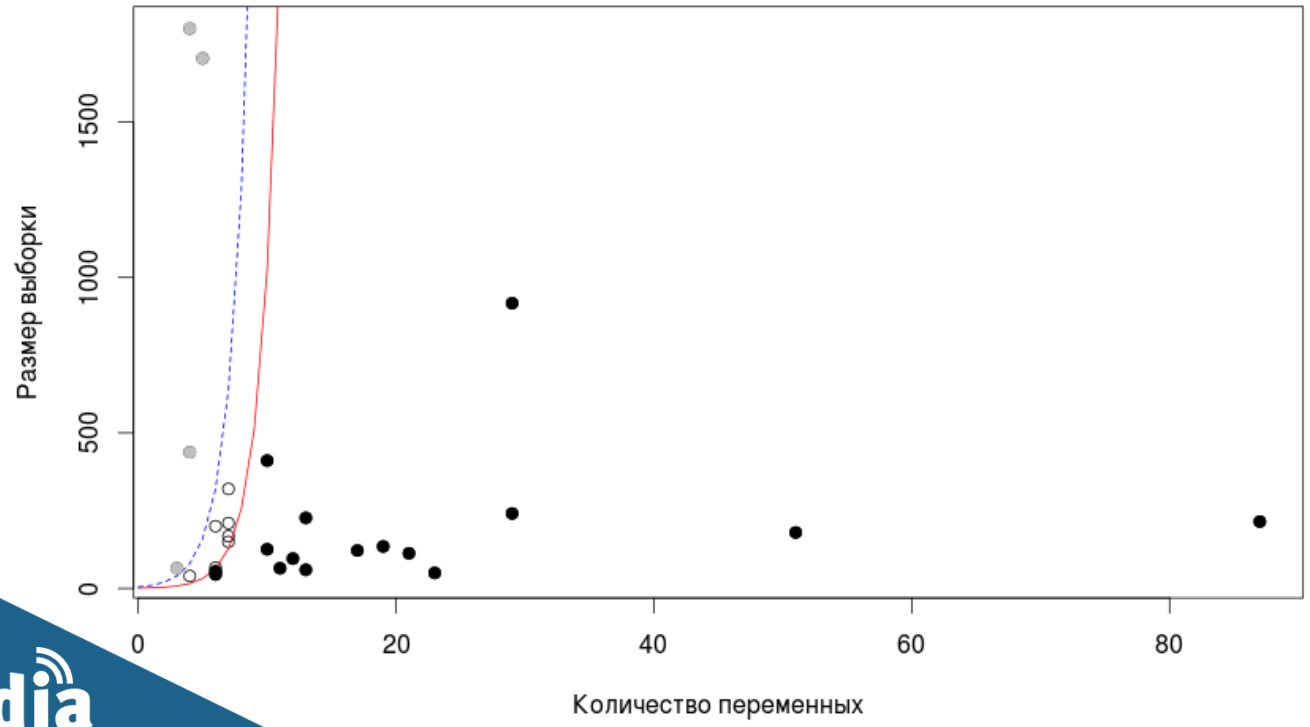
1. Планирование необходимого размера выборки.
2. Выбор числа кластеров.
3. Проведение эксплораторного анализа данных, предварительная подготовка данных при необходимости.
4. Выбор программного обеспечения для вычисления результатов кластерного анализа.
5. Оценка надежности и валидности кластерных решений.



Promedia consulting

# Размер выборки

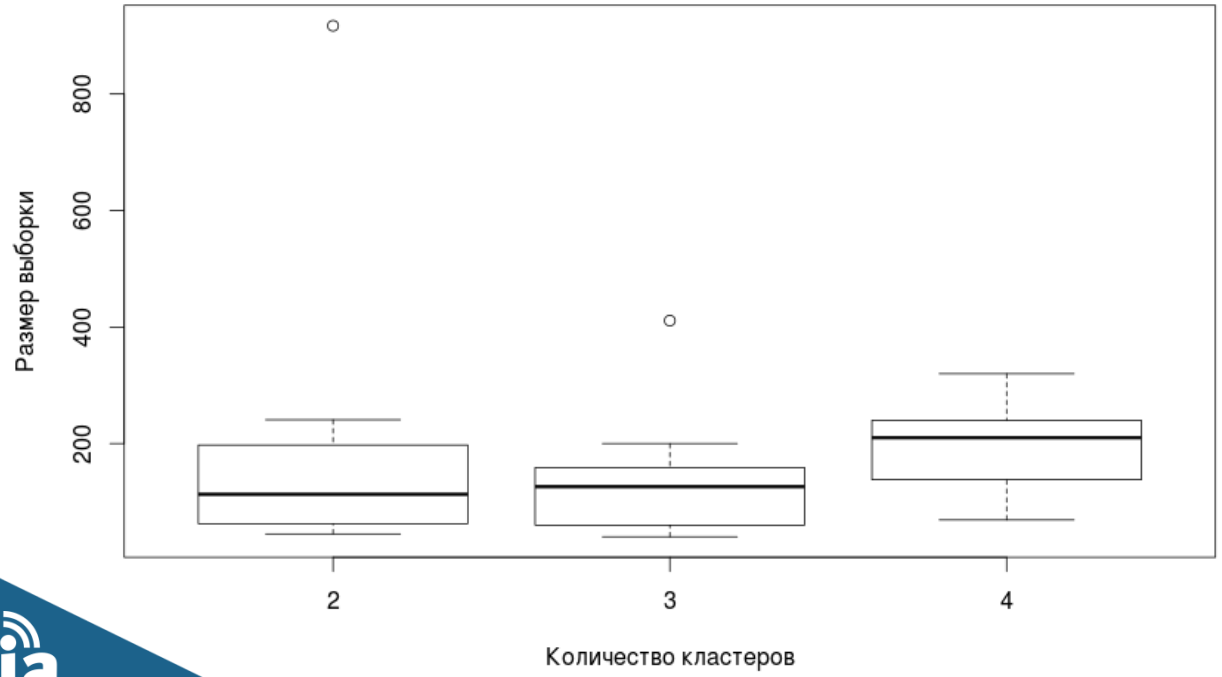
- Не существует общепринятых правил в отношении минимально необходимого размера выборки
- А. К. Форман рекомендует размер выборки не менее 2^k
- Предпочтительный размер выборки должен составлять 5·2^k.



# Число кластеров

Кластеры можно выделить:

- на основе предварительной информации;
- эмпирическим путем;
- визуально определить число кластеров.

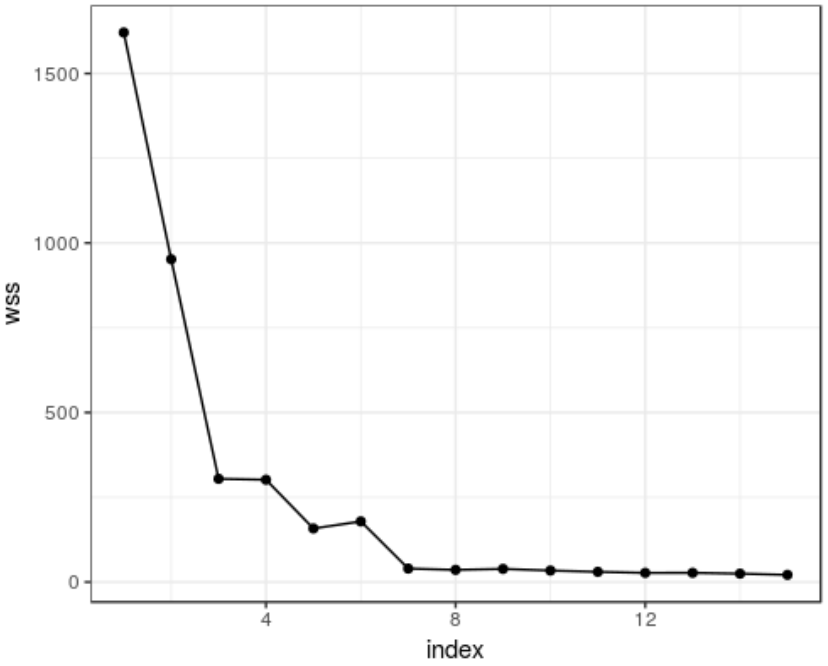



Promedia consulting

# Число кластеров

Критерий «каменистой осыпи»

```
wss <- (nrow(data)-1)*sum(apply(data,2,var))

set.seed(1)

for (i in 2:15)
wss[i] <- sum(kmeans(data, centers = i)$withinss)

ggplot(data.frame(wss = wss, index = 1:length(wss)),
   aes(x = index, y = wss, group = 1)) +
   geom_line() + geom_point() +
   theme_bw()
```



# Число кластеров

Другие варианты в R:

- Проведение иерархического кластерного анализа (**hclust**)
- Использование BIC-критерия (**Mclust**)
- Более 30 различных индексов в пакете **NbClust**

# Предварительное **преобразование данных**

Использование **эвклидова расстояния** имеет смысл, когда:

- наблюдения берутся из генеральных совокупностей, имеющих многомерное нормально распределение, переменные взаимно независимы и имеют равные дисперсии;
- переменные однородны по своему физическому смыслу и одинаково важны для классификации;
- все переменные имеют одинаковые единицы измерения

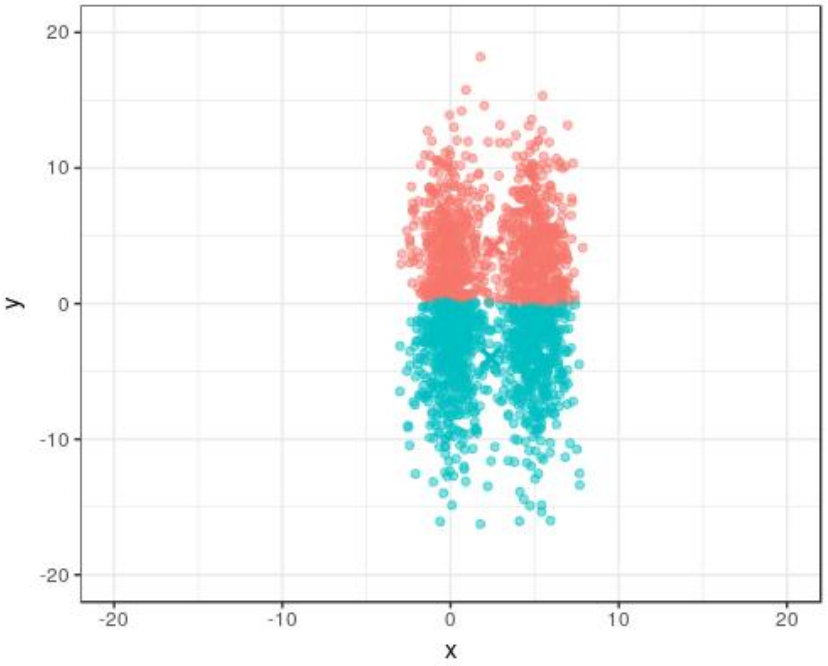
# Предварительное **преобразование данных**

Что следует сделать?

- Принять решение об исключении из объектов **явных выбросов** (в том числе многомерных выбросов) в случае их наличия;
- **Стандартизация** или взвешивание данных для обеспечения «одинаковых» единиц измерения;
- Работа с **коррелированными переменными** — отбор переменных для анализа или применение редукции данных.

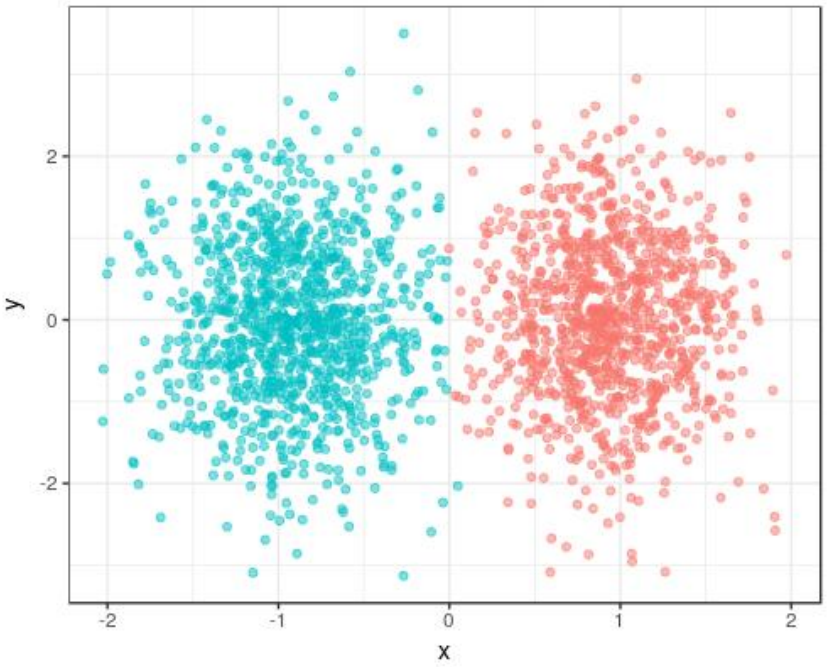
# Предварительное преобразование данных

```
set.seed(1)
data1 <- data.frame(
    x = rnorm(1000), y = rnorm(1000)*5)
data2 <- data.frame(
    x = rnorm(1000) + 5, y = rnorm(1000)*5)

data <- as.data.frame(rbind(data1, data2))

set.seed(1)
clusters <- kmeans(data, centers = 2)
```



# Предварительное **преобразование данных**

```
data <- scale(data)

set.seed(1)
clusters <- kmeans(data, centers = 2)
```




Promedia consulting

# Предварительное **преобразование данных**

Почему следует быть осторожными с **кластерами на факторах?**

- Нет никаких гарантий, что в следующем наборе данных повторится полученная структура компонент;
- Происходит потеря исходной дисперсии данных, что может вести к искажению решения;
- Устранение групп переменных с малыми нагрузками может привести к потере важной информации о нишевых сегментах данных;
- интерпретация кластерных решений может быть затруднена

# Выбор **начальных центров**

Для избежания **неоптимальной кластеризации** возможны такие варианты:

- назначать центры на основе существующих знаний и теорий
- использовать центры иерархической кластеризации
- ограничивать выбор начальных центров участками с высокой плотностью данных
- «переопределять» начальные центры при помощи бутстрепа;
- провести большое количество случайных разбиений (более 5000) на K групп и выбрать итоговое решение (nstart)

# Выбор **программного обеспечения**

При написании работ крайне важно точно **указывать программное обеспечение**, которое вы использовали для получения кластерного решения!

Promedia consulting

# Выбор **программного обеспечения**


По умолчанию использует алгоритм Ллойда. Использует K первых наблюдения в качестве начальных центров. Есть алгоритм МакКуина и возможность задать центры


По умолчанию использует алгоритм Ллойда. Использует K первых наблюдения в качестве начальных центров.


Имеет функции FASTCLUST (МакКуин) и PROC FASTCLUST (Хартиган-Вонг). Центры выбираются по мере уменьшения плотности данных.

# Выбор **программного обеспечения**


Имеет функцию FindClusters (Ллойд/Форджи). Использует метод K-медоиды (центр кластера является объектом набора данных)


Использует двухфазный алгоритм – вначале используются частичные данные, а только затем итеративный процесс


Применяется метод Ллойда, а начальные центры выявляются при помощи алгоритма “kmeans++”


Команда kmeans по умолчанию использует алгоритм Хартигана-Вонга. Начальные центры выбираются случайным образом.

# Оценка **кластерного решения**

**Надежность**


Необходимо провести несколько однотипных исследований или, как минимум, разбить исходный массив данных на несколько частей и провести кластеризацию отдельно.

Использовать другие алгоритмы для кластеризации данных. Сопоставить результаты.

# Оценка **кластерного решения**

**Валидность**

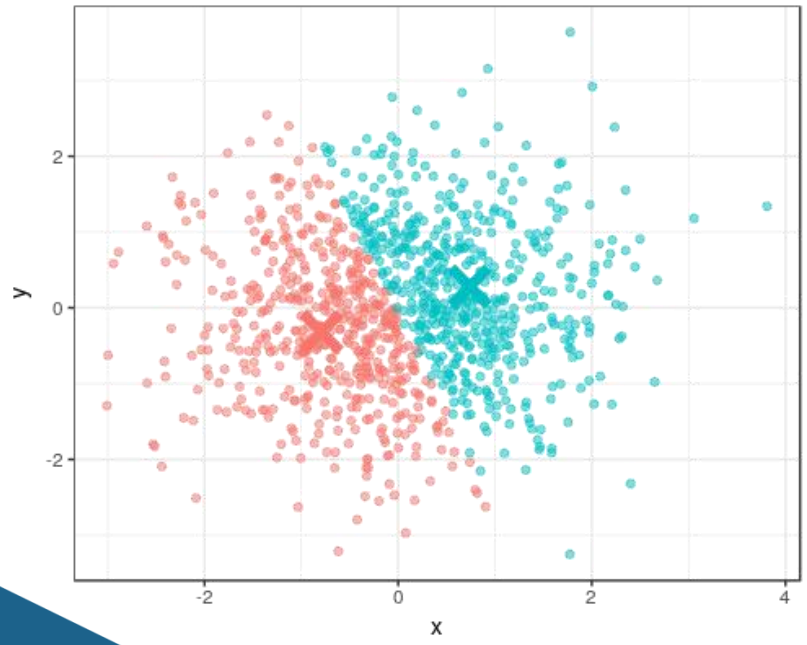
Для оценки валидности существуют вненшие и внутренние критерии.

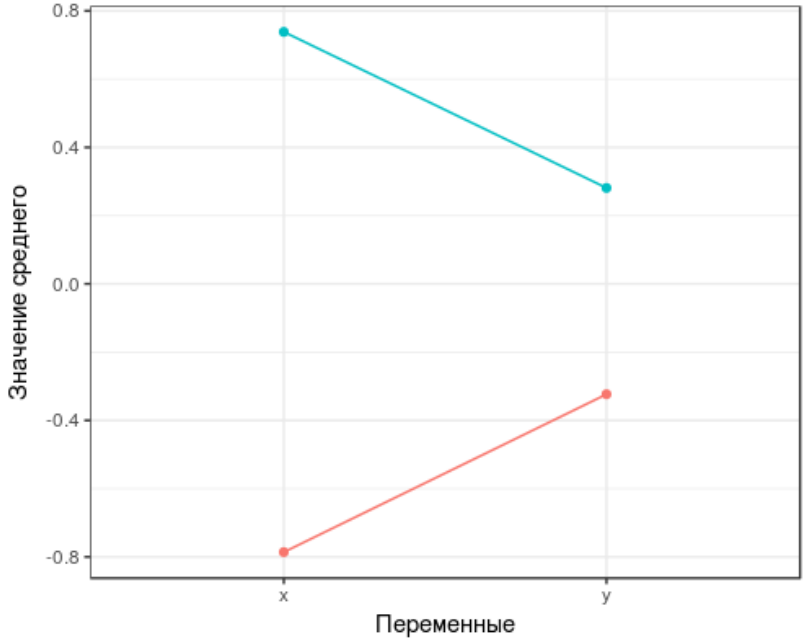
Внешняя валидность оценивается при помощи экспертных оценок или же переменных, которые не использовались при кластеризации.

Внутренняя валидность использует для проверки статистические процедуры над переменными, на основе которых производилась кластеризация. Тем не менее, следует быть осторожными в интерпретации кластерных структур.

# Оценка **кластерного решения**

Простое применение критериев различий для оценки внутренней валидности недостаточно. Необходимо использовать процедуры многомерного шкалирования, бутстрепа, дискриминантного анализа и аналогичных процедур.

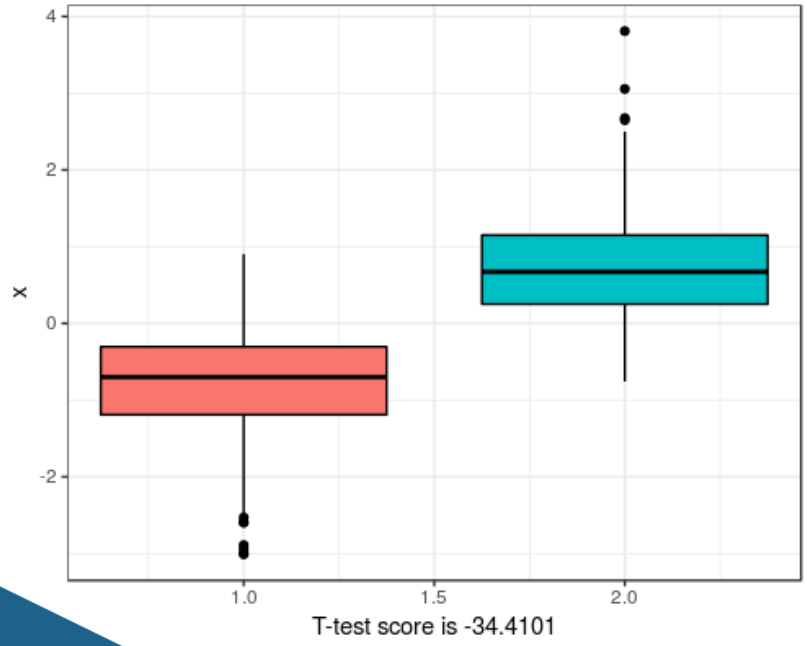


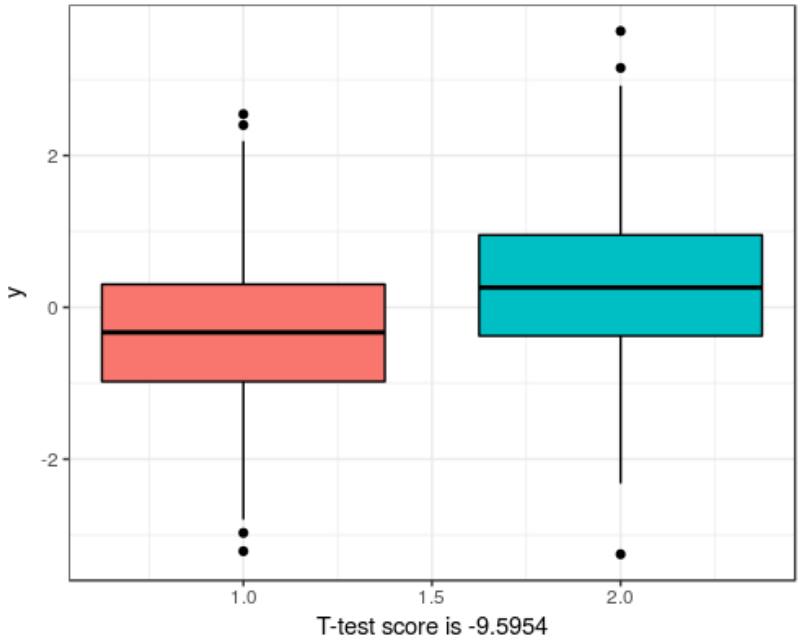




Promedia consulting

# Оценка **кластерного решения**

Простое применение критериев различий для оценки внутренней валидности недостаточно. Необходимо использовать процедуры многомерного шкалирования, бутстрепа, дискриминантного анализа и аналогичных процедур.




Promedia consulting


K-means
Сейчас можно задать вопросы
Promedia
consulting