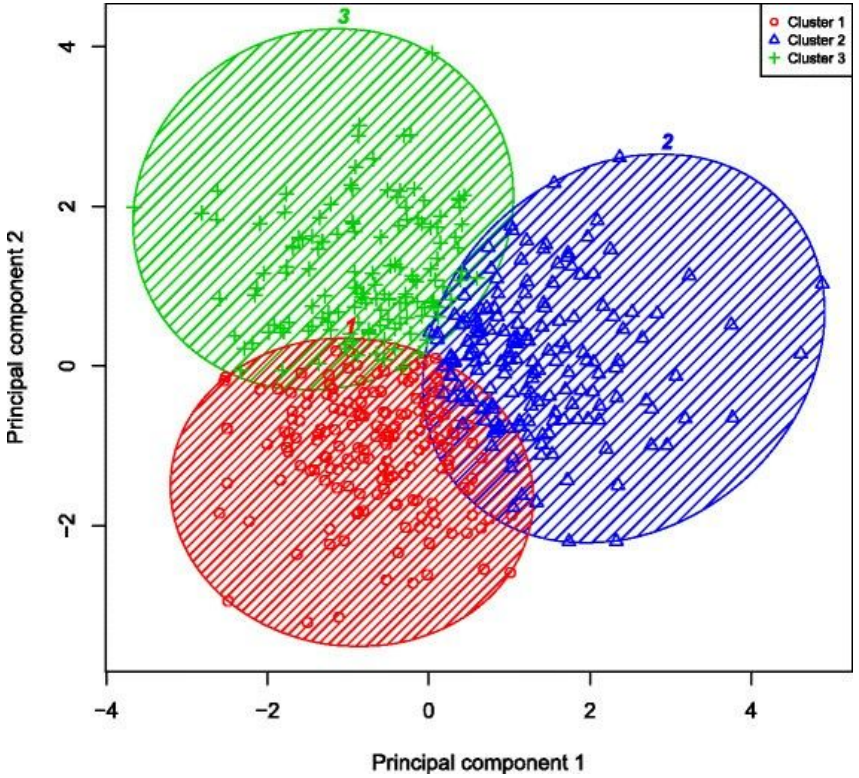
Кластерный анализ

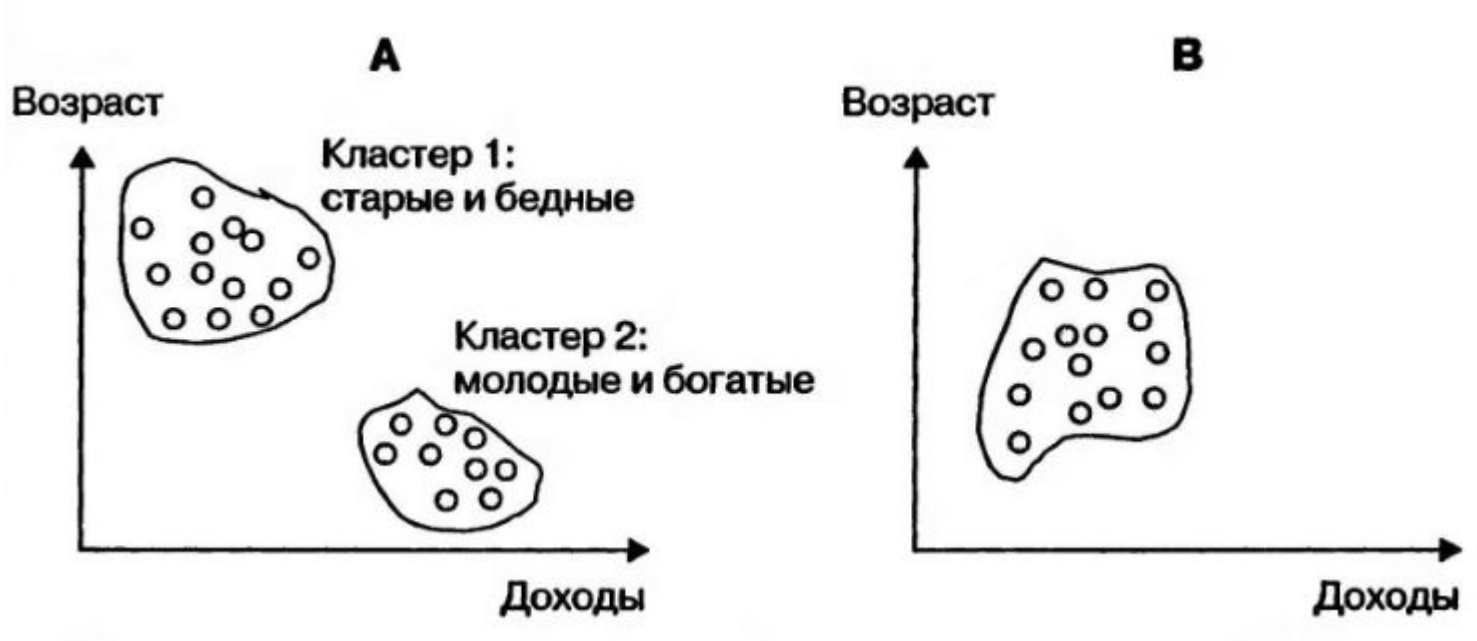
Кластерный анализ - многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы



Применение кластерного анализа

1. в маркетинге — для сегментирования клиентов, конкурентов, исследования рынка
2. в медицине — для кластеризации симптомов, заболеваний, препаратов
3. в биологии — для классификации животных и растений
4. компьютерных науках — для группировки результатов при поиске сайтов, файлов и других объектов

Курильщики сигар, возраст и уровень доходов которых известны, исследуются на предмет возможности их разделения на однородные группы

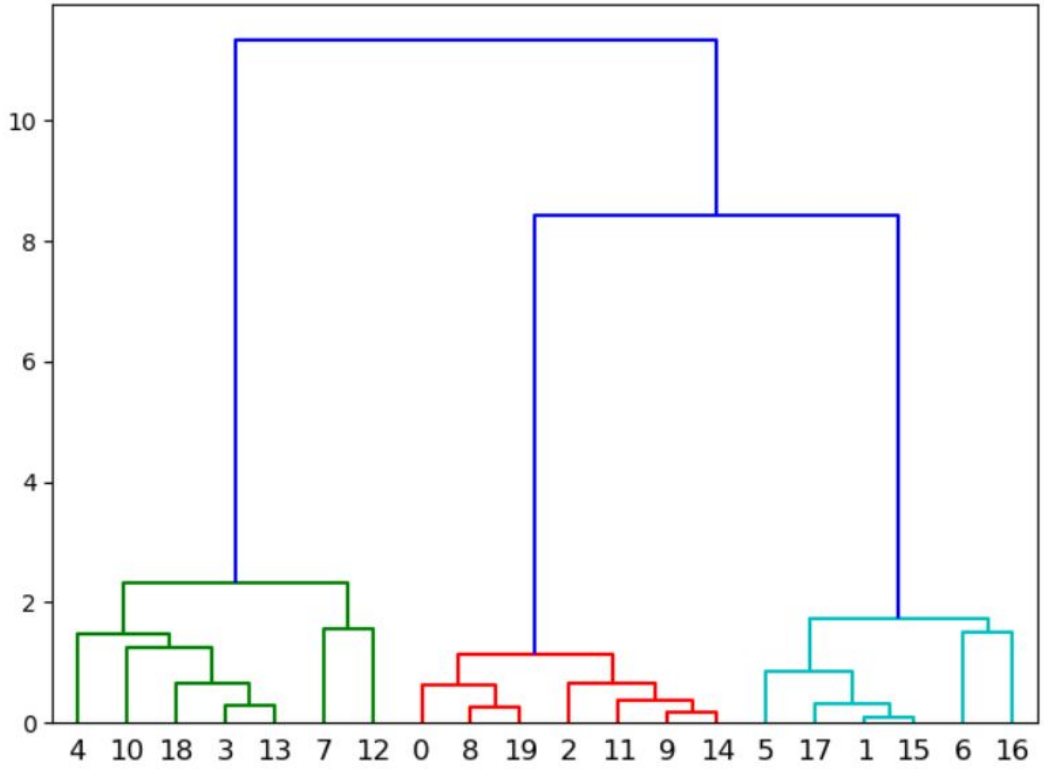


Методы кластерного анализа

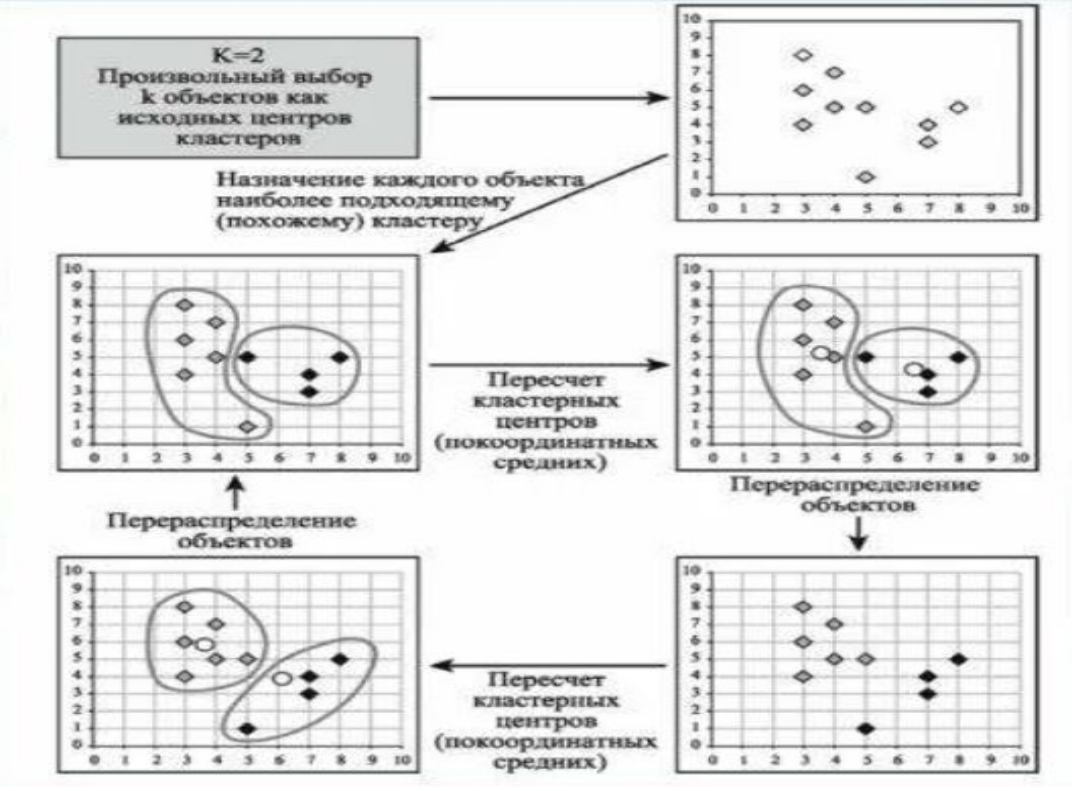
1. Иерархические - первоначально все объекты рассматриваются как отдельные кластеры. Выстраивается дерево кластеров путем объединения первоначальных существовавших кластеров.

2. Итеративные - разбиение на кластеры получается из некоторого начального разбиения способом последовательных итераций. Число конечных кластеров пользователь задает самостоятельно.

Дендрограмма - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров



Метод k-средних - это алгоритм, смысл которого заключается в наблюдении за набором немаркированных данных для автоматического обнаружения скрытой структуры, а также для обнаружения закономерности в немаркированных данных.



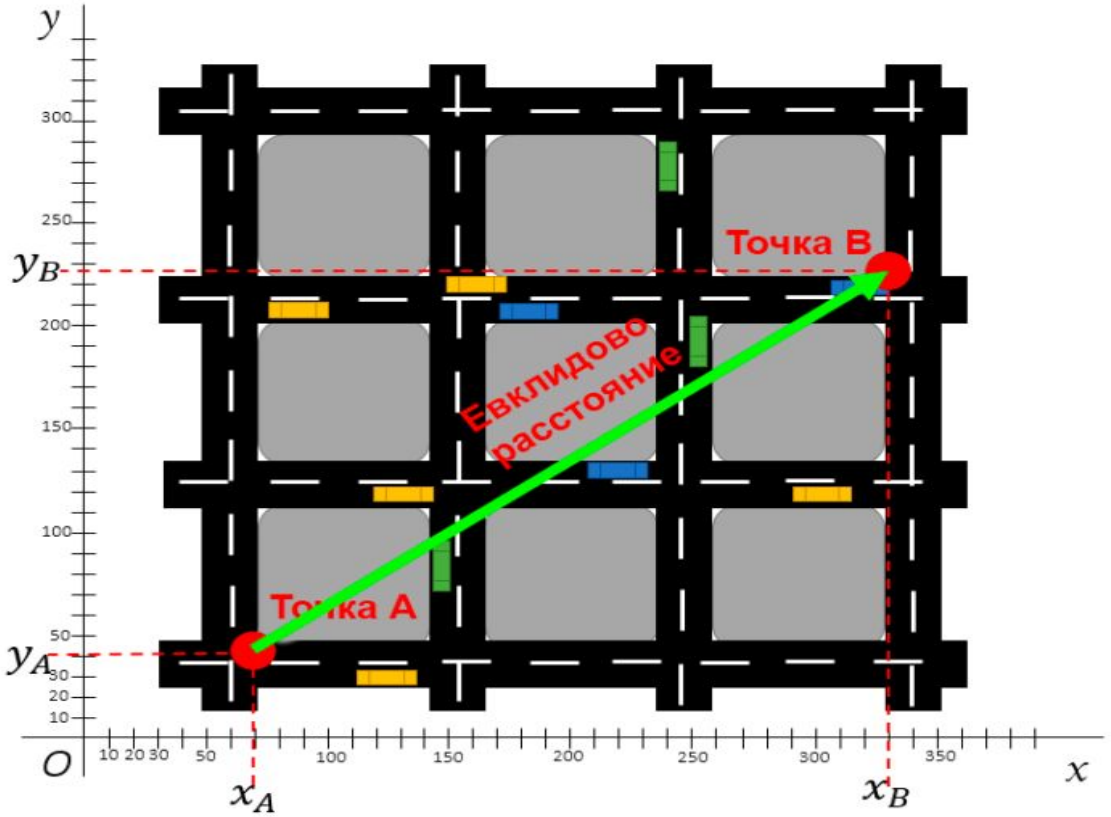
Метрики расстояний

1. **Евклидово расстояние — это прямая линия между двумя точками с координатами X и Y (кратчайший путь).**

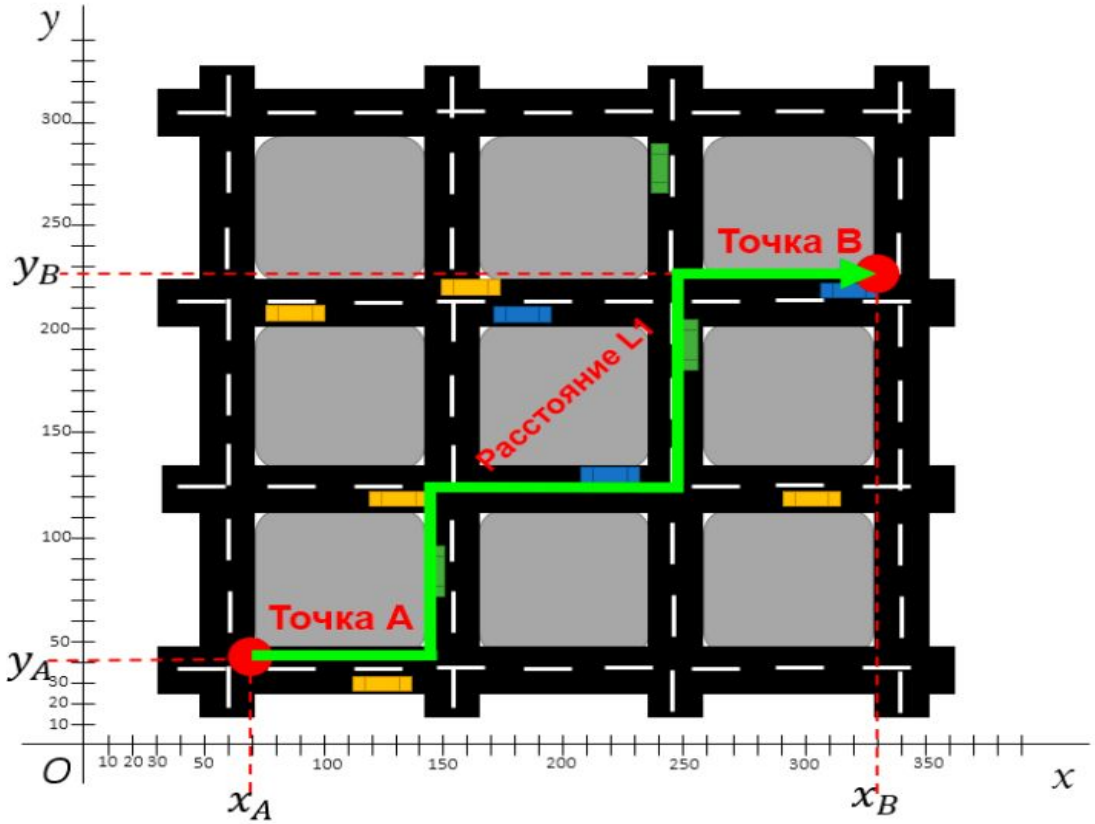
2. **Манхэттенское расстояние (L1)**— измеряет дистанцию не по кратчайшей прямой, а по блокам. Расстояние L1 измеряет дистанцию между городскими блоками: это расстояние всех прямых линий пути.

3. **Расстояние Чебышева — метрика на векторном пространстве, задаваемая как максимум модуля разности компонент векторов.**

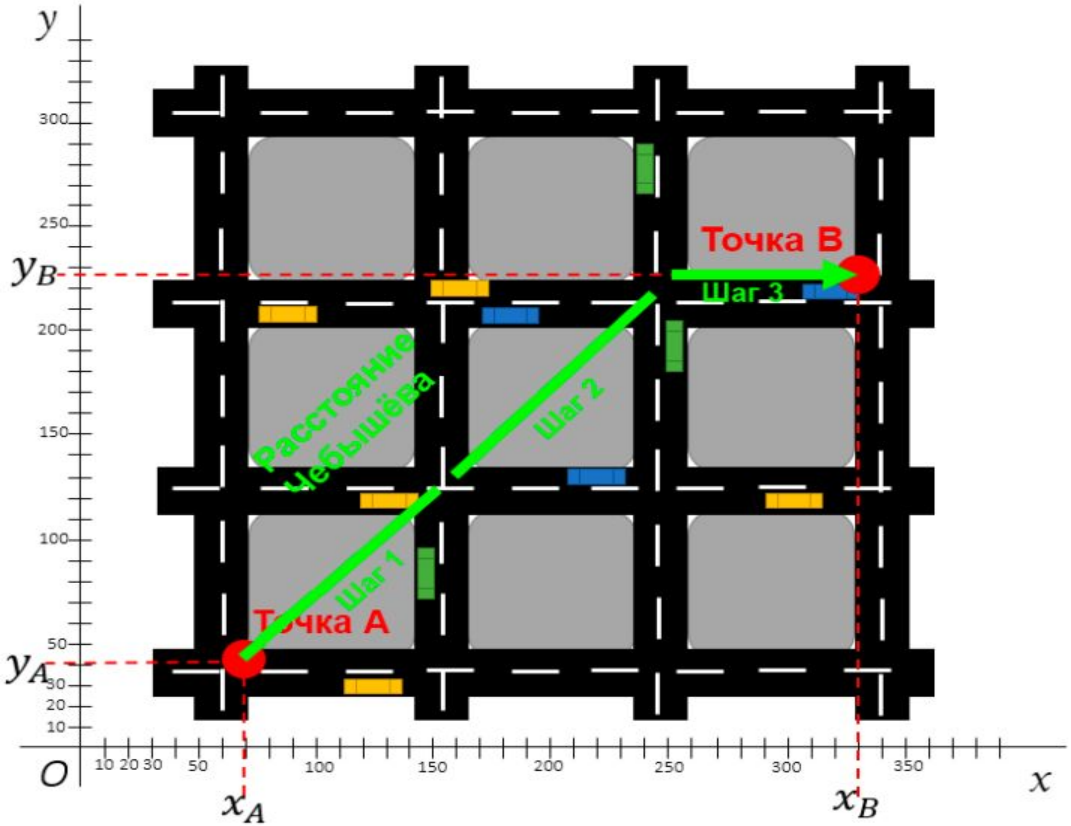
Евклидово расстояние



Манхэттенское расстояние



Расстояние Чебышева



В заключение

В отличие от многих других статистических процедур, методы кластерного анализа используются в большинстве случаев тогда, когда вы не имеете каких-либо априорных гипотез относительно классов, но все еще находитесь в описательной стадии исследования.

Спасибо!