**Лабораторная работа №1**

Описание датасета

Данные содержат информацию о 400 покупателях.

Атрибуты:

1. Пол – мужской, женский
2. Возраст от 18 до 70
3. Годовой доход в тысячах $
4. Расходы в процентах 1-100

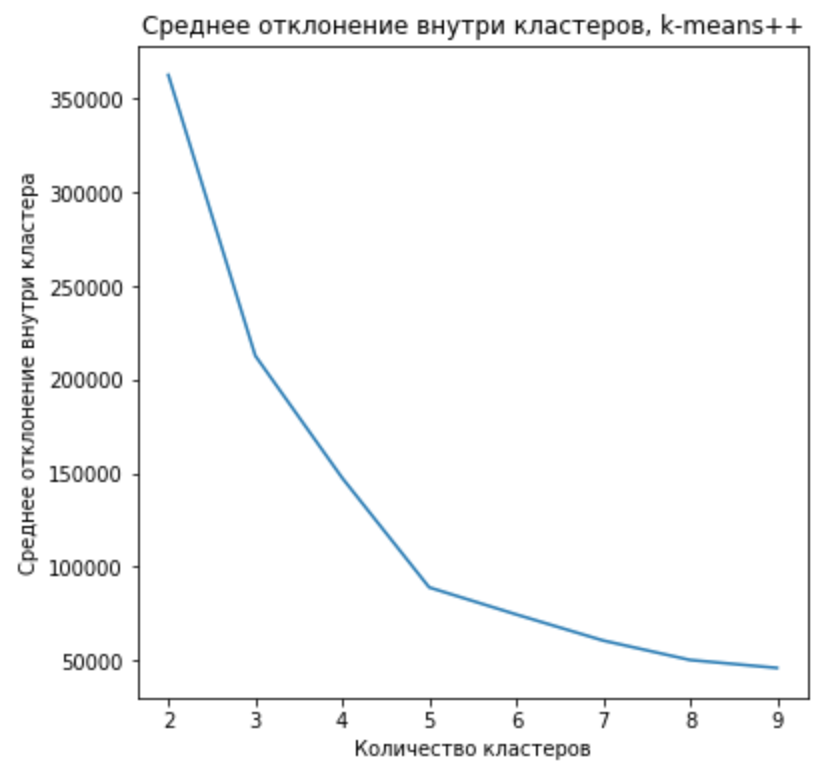
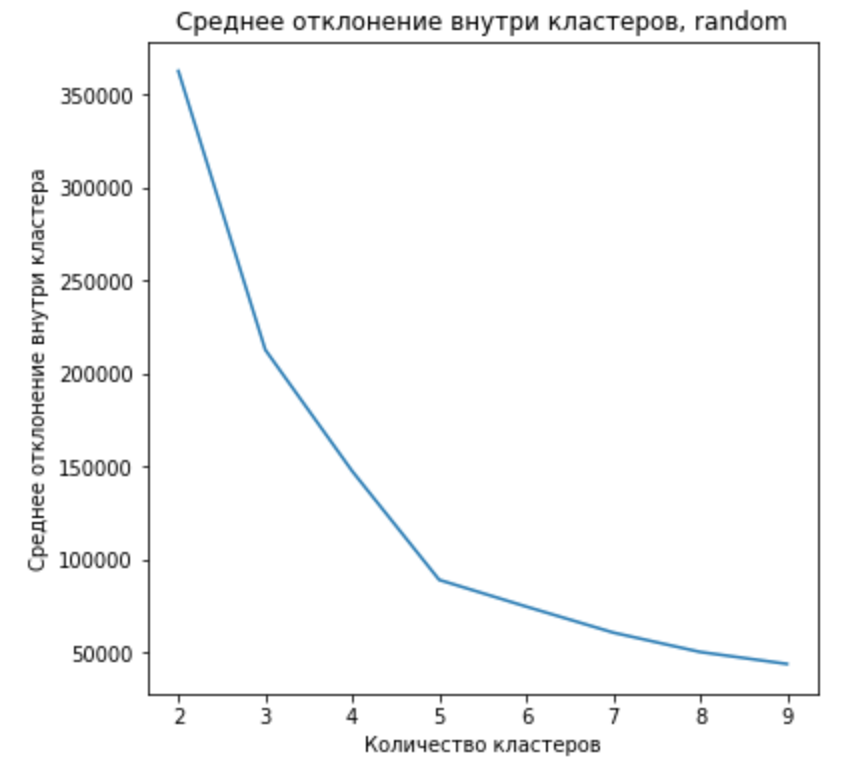
Данные оказались целостными и без пропусков, поэтому дополнительных обработок не потребовалось.



Кластеризация К-средних

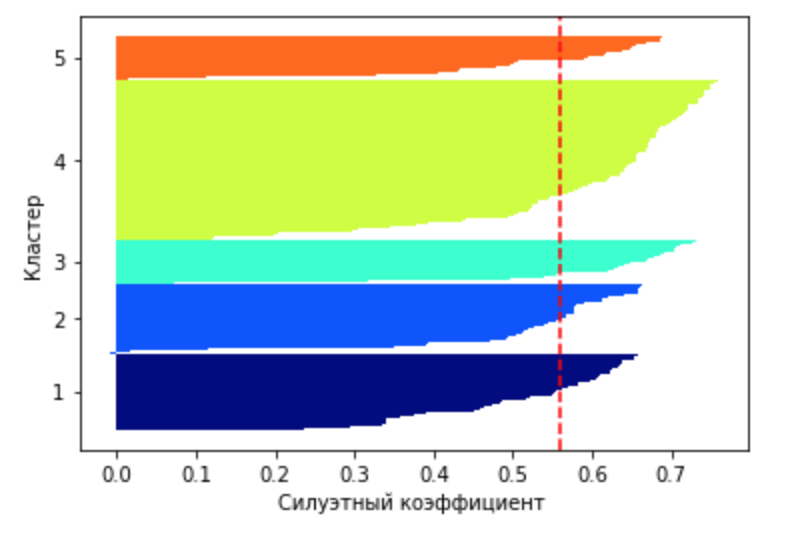
Для кластеризации будем использовать библиотеку sklearn. Библиотечная реализация k-means по-умолчанию выполняет алгоритм несколько раз с различными центрами, выбирая в качестве конечного результата кластеры с наименьшим внутрикластерным отклонением. Число повторений можно переопределить параметром n\_init, по-умолчанию он равен 10. Способ выбора центроидов задаётся параметром init. По-умолчанию используется улучшенный метод k-means++, но сначала воспользуется случайным выбором.

Выполним кластеризацию с разным желаемым количеством кластеров и посмотрим на изменение среднего внутрикластерного отклонения.



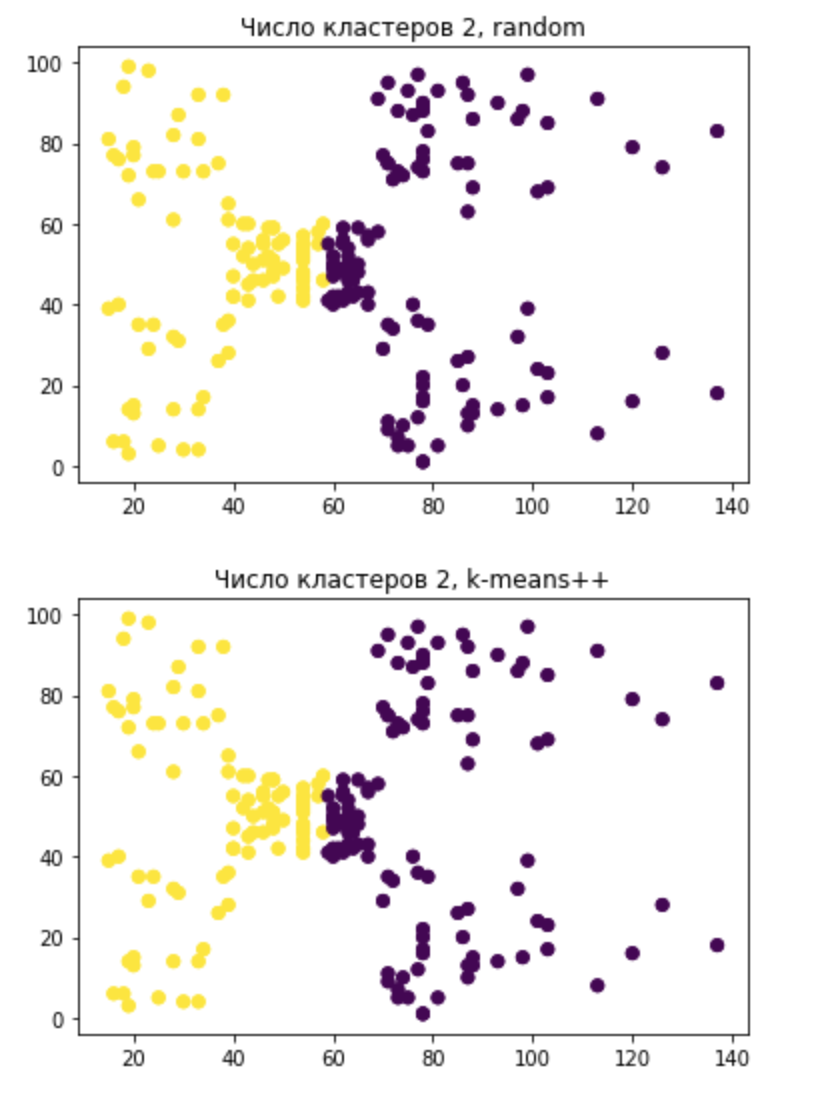
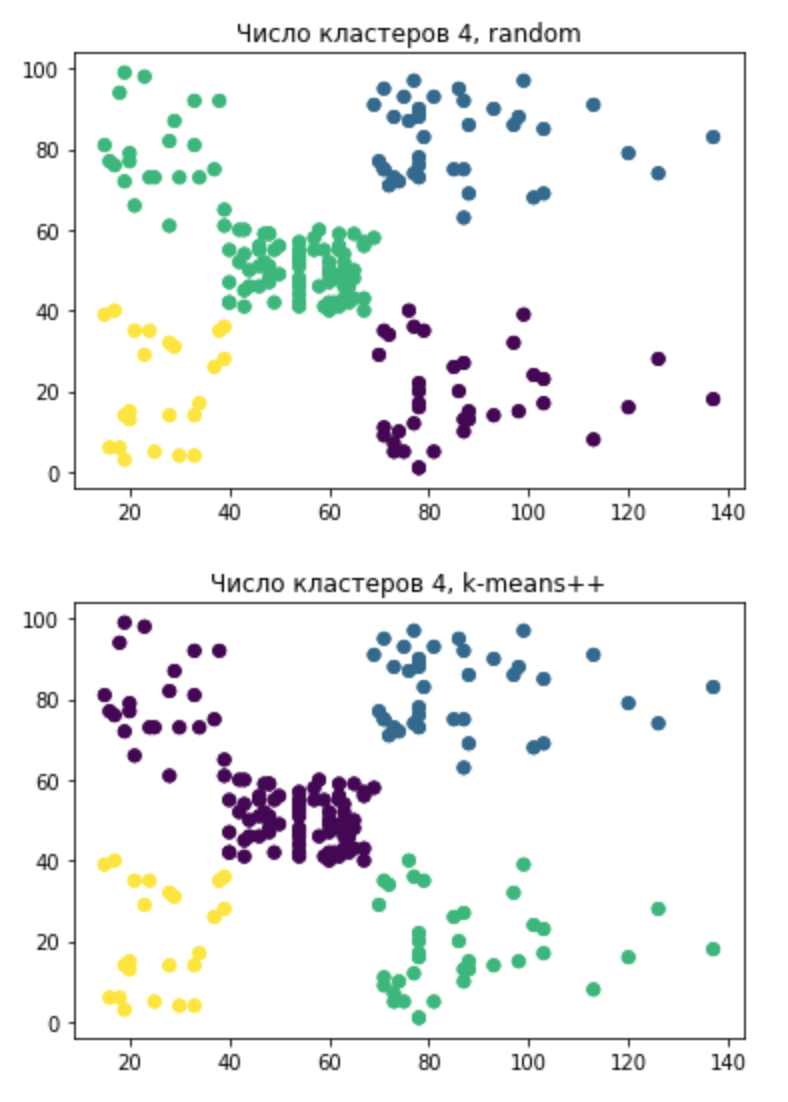
На графике видим, что скорость уменьшения отклонения падает после 5 кластеров. Оба графика примерно одинаковы, поэтому будем использовать около 5 центроидов.

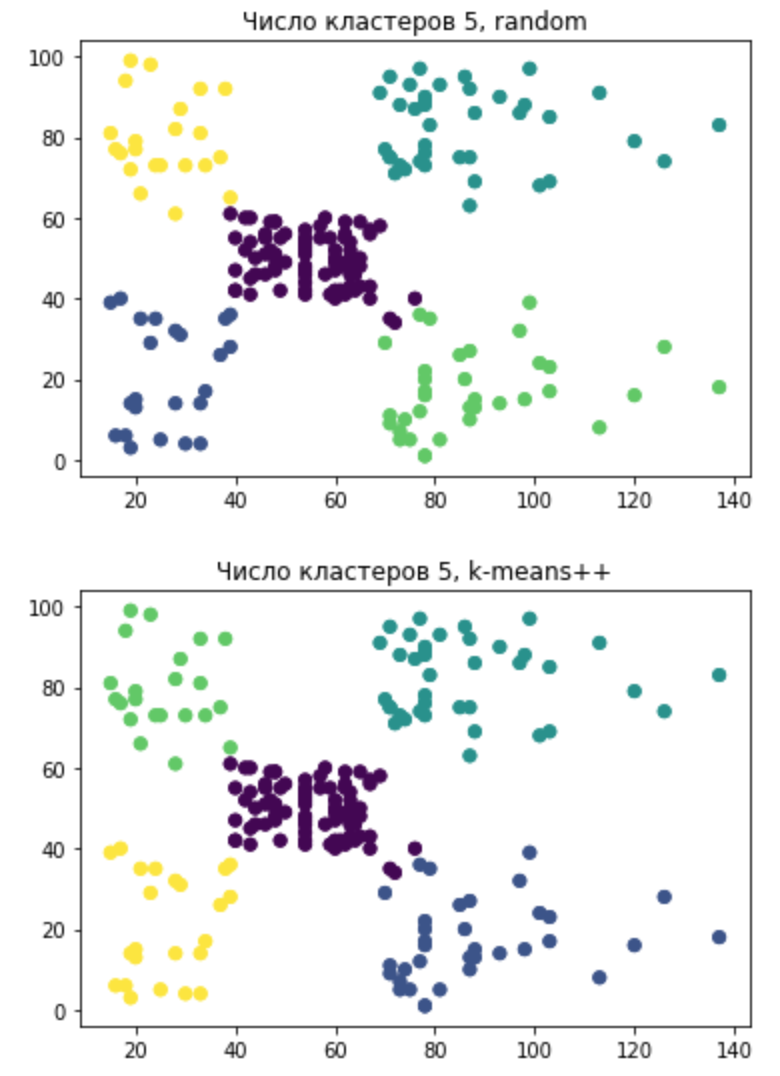
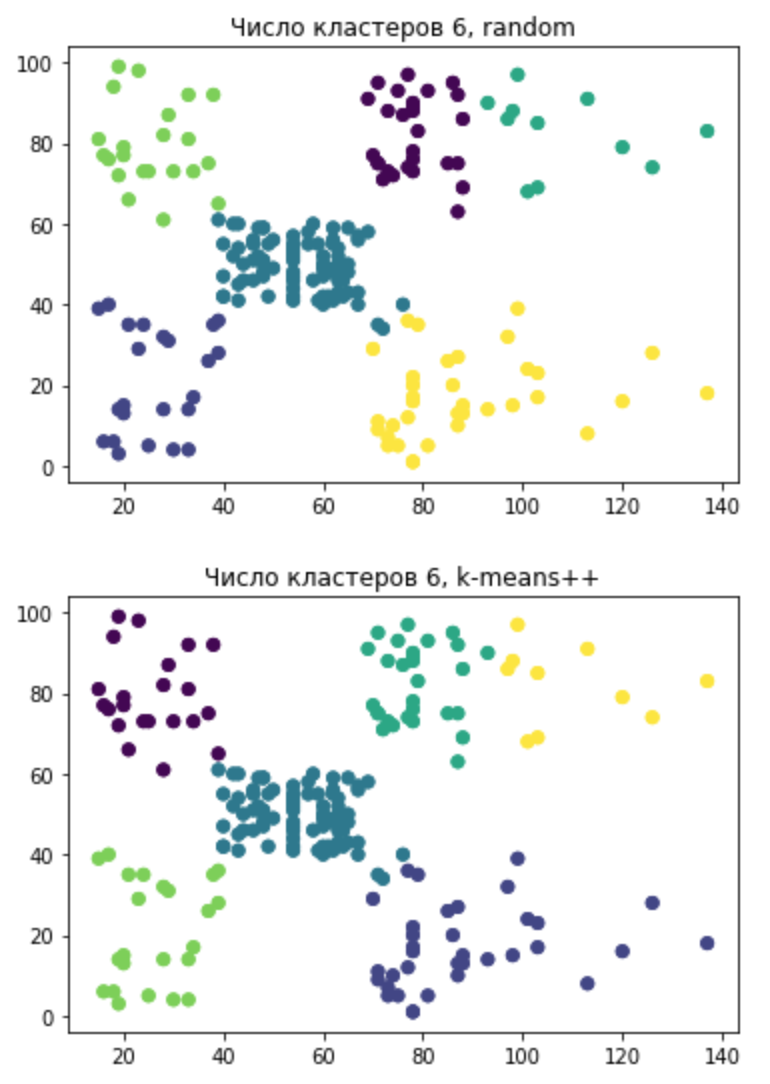
Визуальный анализ силуэтного графика позволяет быстро рассмотреть размеры разных кластеров и идентифицировать кластеры, которые содержат выбросы:



Как видно на приведенном выше силуэтном графике, наши силуэтные коэффициенты даже близко не находятся рядом с О, что может служить индикатором хорошего объединения в кластеры. Более того, чтобы резюмировать качество нашей кластеризации, мы добавили в график средний силуэтный коэффициент (пунктирная линия).

Оценим визуально кластеры, получившиеся при разном параметре.

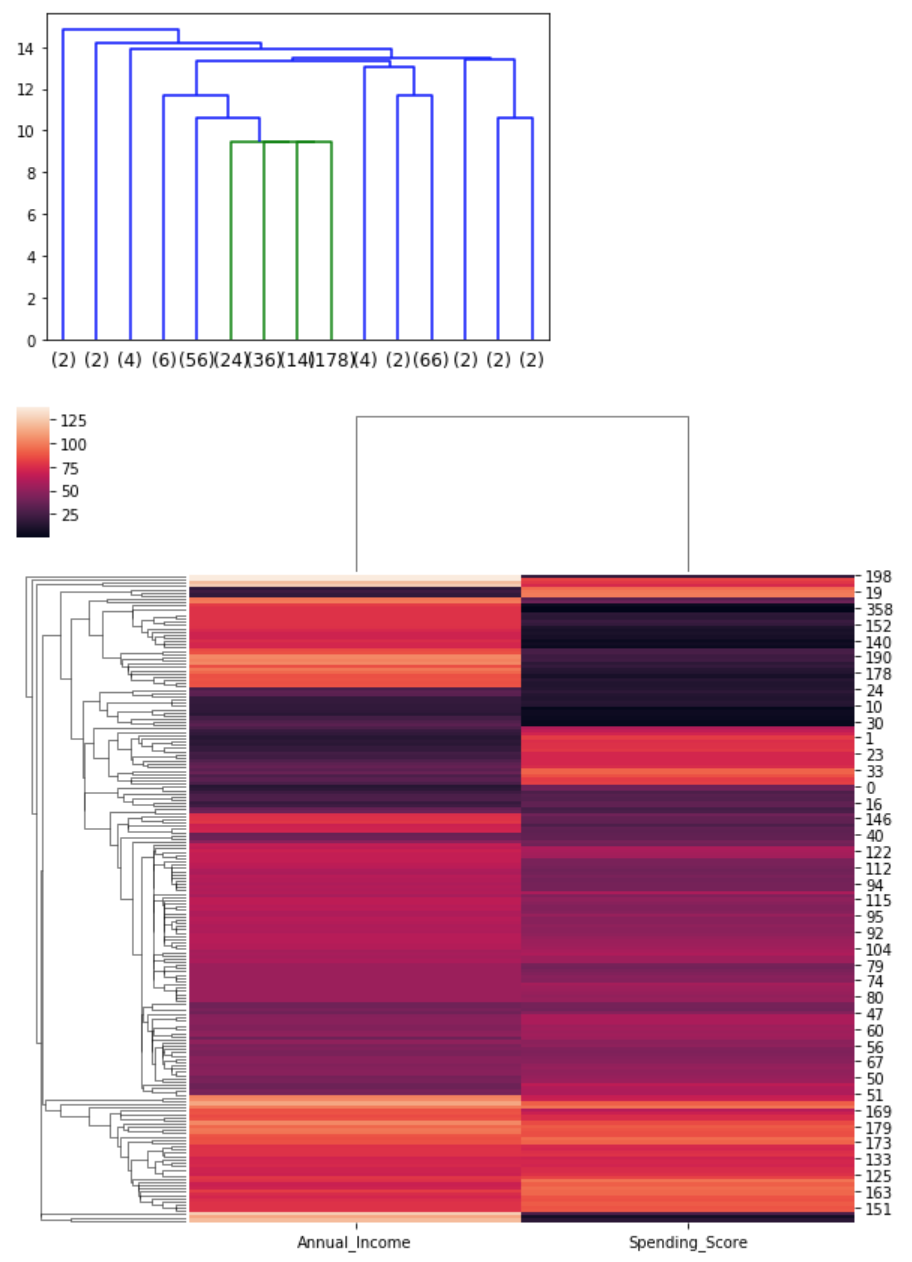
 

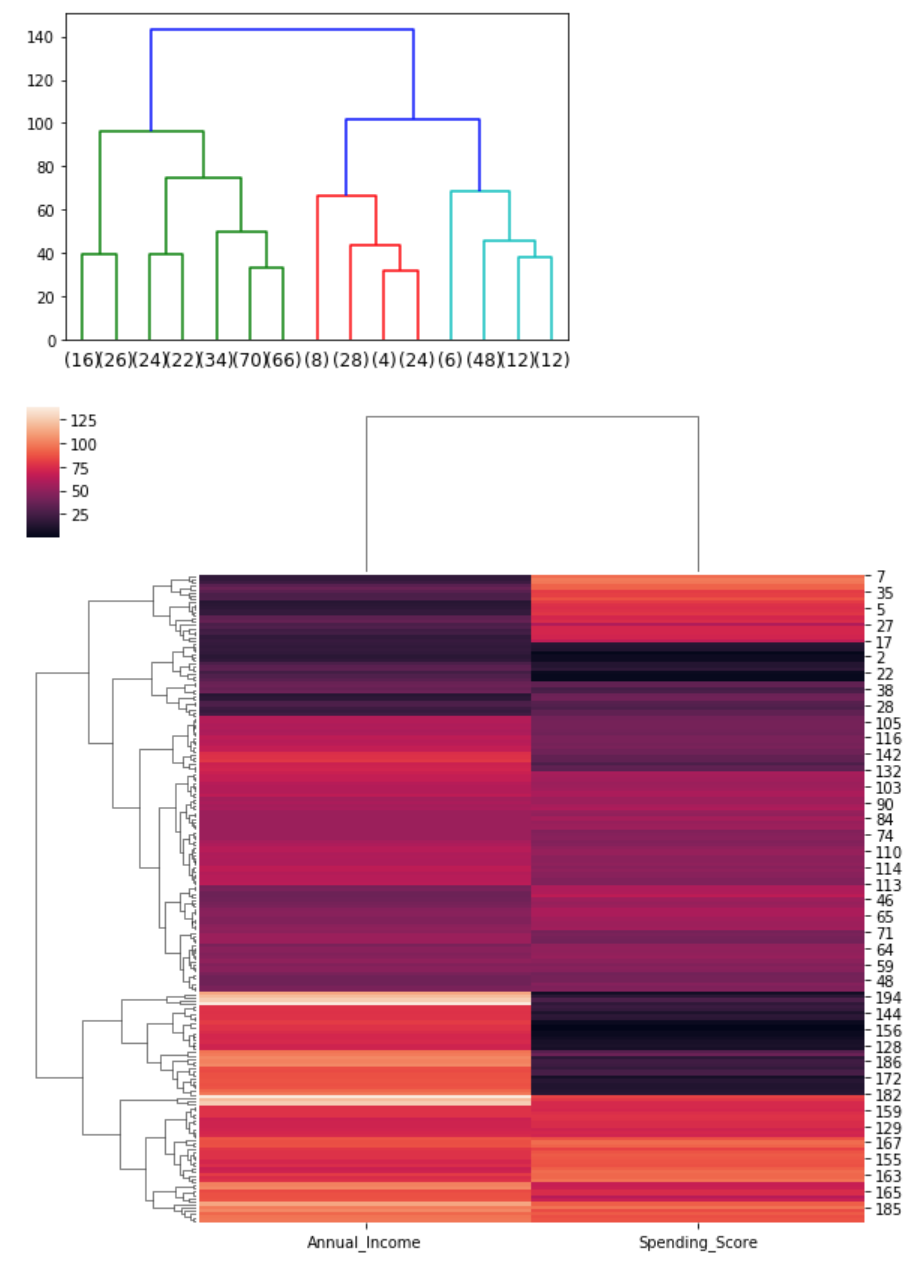
При кластеризации на 5 кластеров оба метода выбора центроидов показали себя отлично.

Иерархическая кластеризация

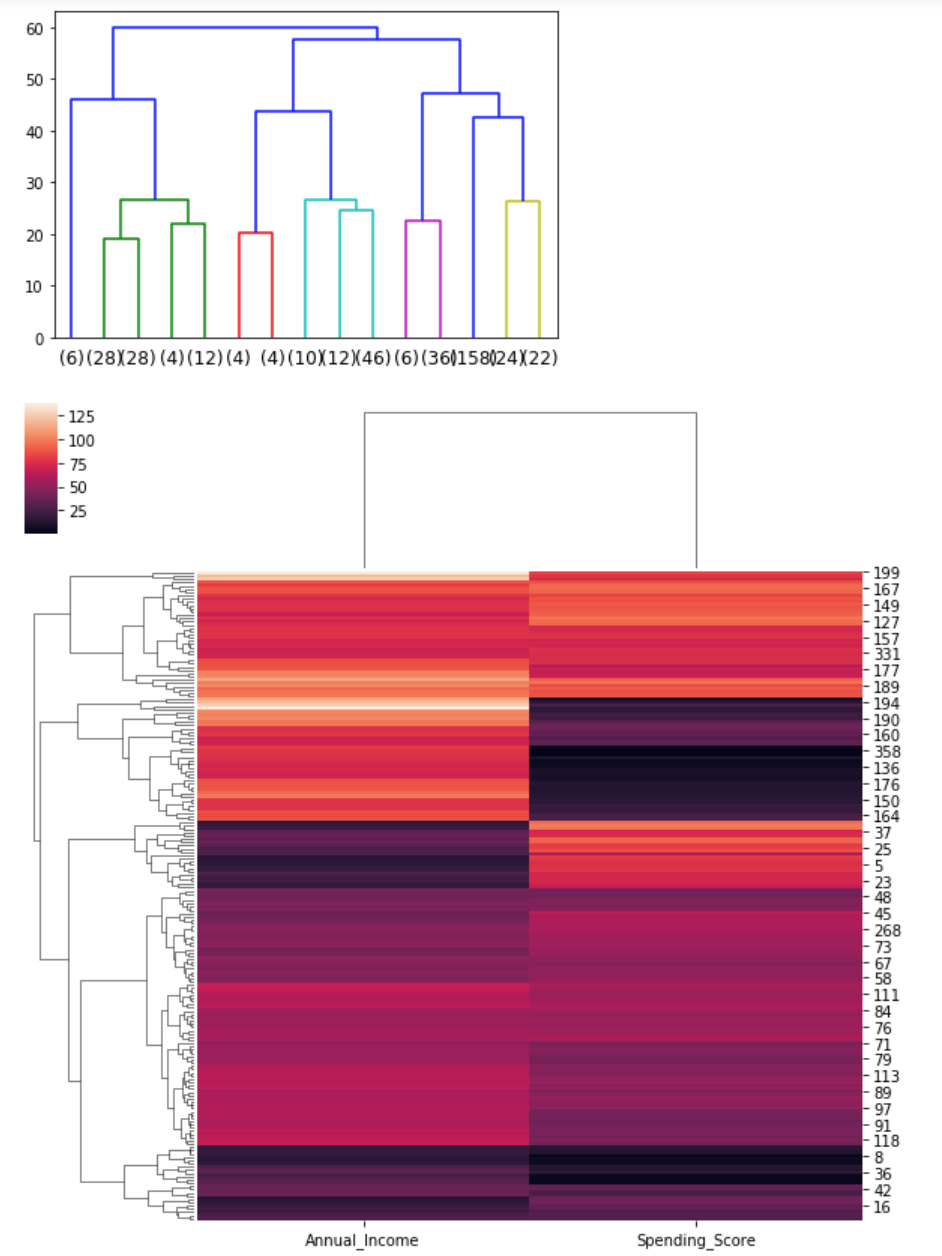
Для кластеризации будем использовать библиотеку scipy и seaborn для тепловой карты. Ниже приведено сравнение разбиения на кластеры при разных способах измерения расстояния между кластерами - single, complete, average, ward.



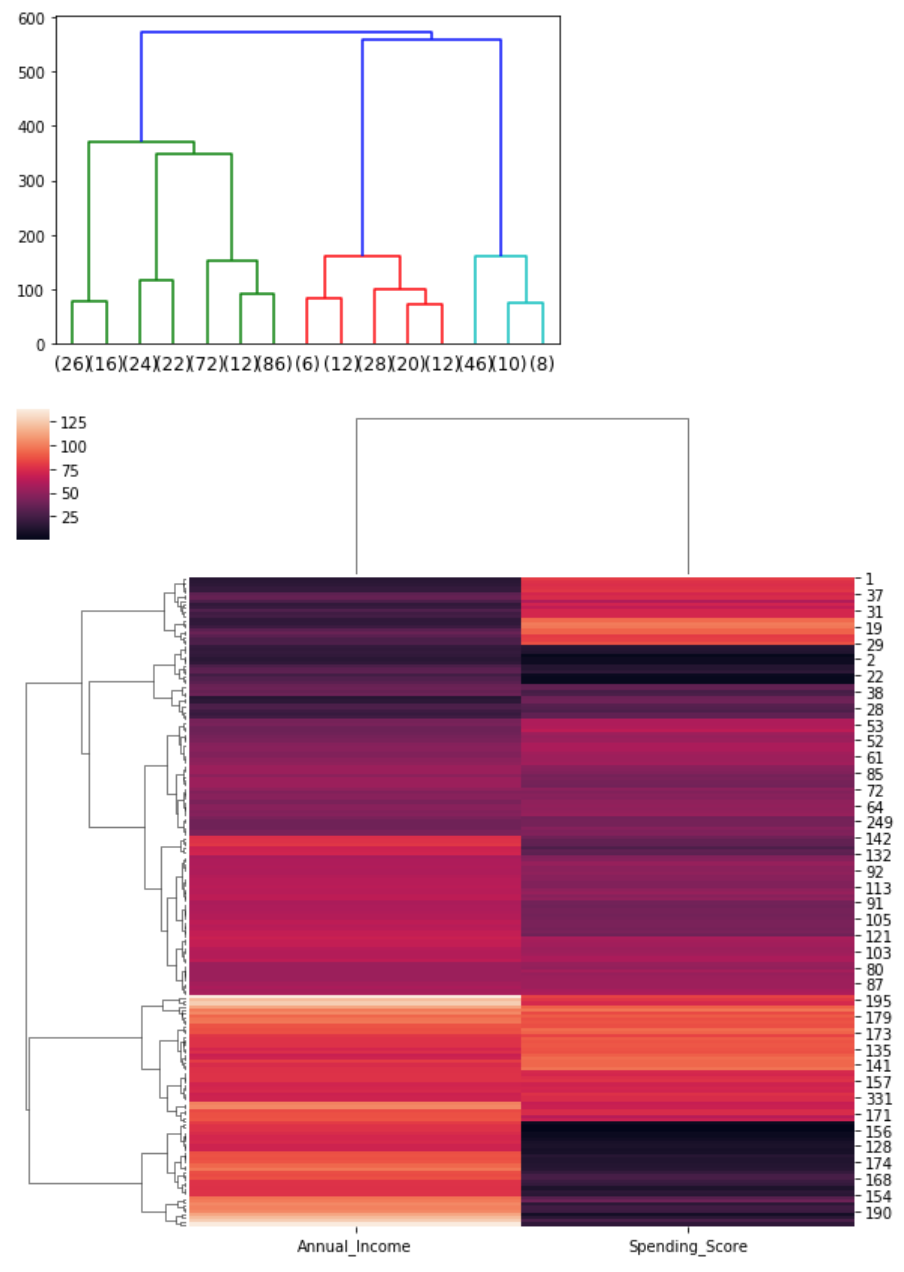
Single - расстояние между кластерами как расстояние между самыми близкими точками из разных кластеров. При таком методе алгоритм видит два ярковыраженных кластера и много похожих при более мелкой детализации.



сomplete - расстояние между кластерами как расстояние между самыми отдаленными точками из разных кластеров



average - среднее расстояние по декартовому произведению точек двух кластеров



ward - иcпользуется метод Уорда

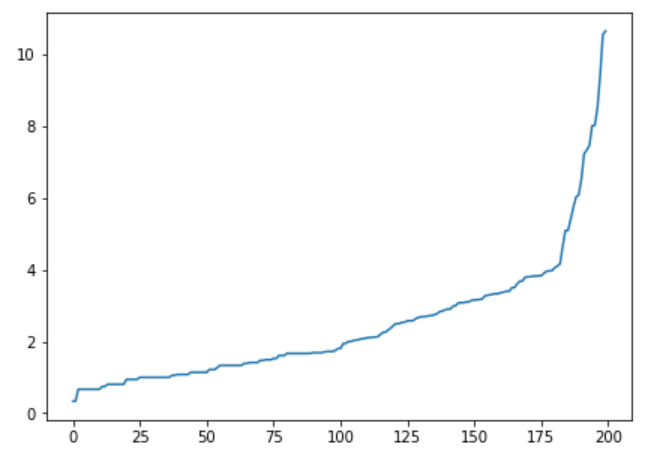
Кластеризация DBSCAN

Для кластеризации методом DBSCAN воспользуемся реализацией из библиотеки sklearn. Модель предоставляет 2 основных параметра:

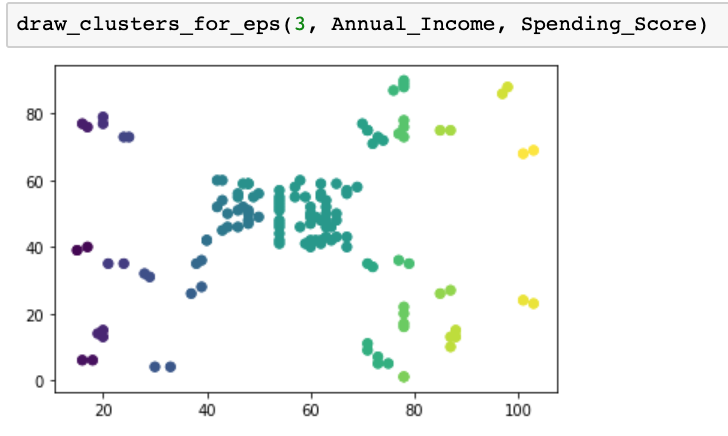
eps - максимальное расстояние между двумя точками, при котором они распознаются как соседи

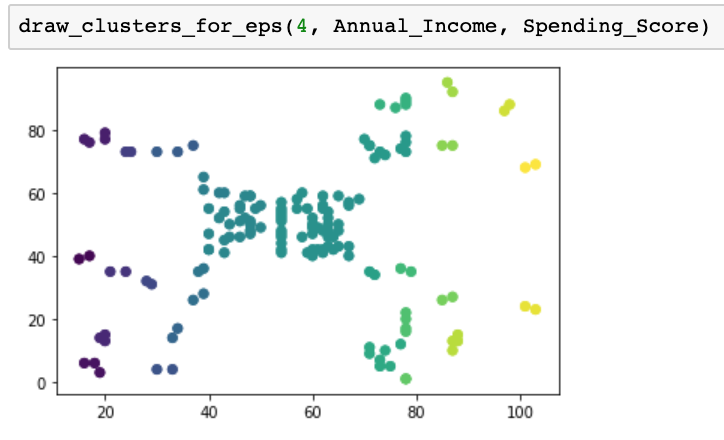
min\_samples - минимальное количество точек, которые могут образовать кластер, возьмём минимальное значение - 3

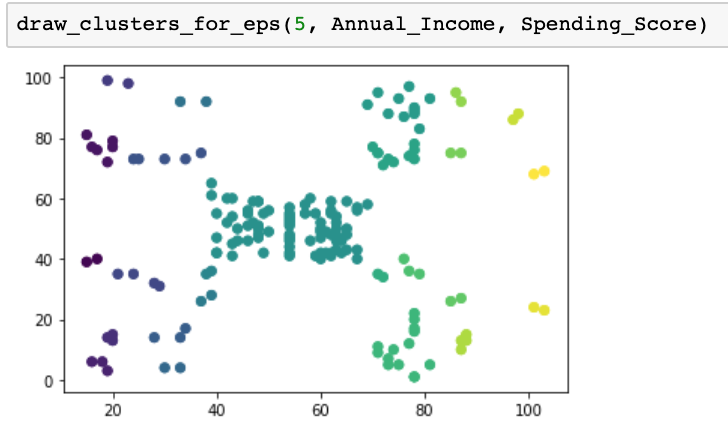
Посмотрим, как изменяется среднее попарное расстояние между точками при различном eps.

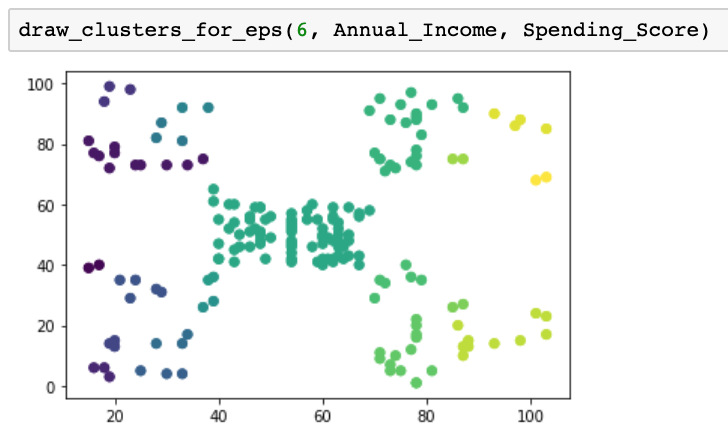


Видим, что большинство точек имеет среднее расстояние < 4, после чего наблюдается резкий скачок. Рассмотрим полученные кластеры при eps = (3, 4, 5, 6). Алгоритм помечает специальным лейблом -1 точки, которые он посчитал шумом. Их будем убирать с графика.









**Вывод**

На практике были изучены методы преобразования данных, а так же три алгоритма кластеризации.

Лучше всего на взятых данных себя показал алгоритм К-средних при 5 центроидах. Учитывая, что кластеризация является методом обучения без учителя, мы не можем себе позволить роскошь раз и навсегда определенных данных о метках классов с целью оценки качества модели. Поэтому мы обратились к полезным внутренним метрикам оценки качества, таким как метод локтя или силуэтный анализ, в качестве попытки количественно определить качество кластеризации.

Затем мы проанализировали другой подход к кластеризации: агломеративную иерархическую кластеризацию. Иерархическая кластеризация не требует предварительно указывать число кластеров, и ее результат может визуализироваться в виде дендограммы, которая помогает интерпретировать результаты. Последним алгоритмом кластеризации, который мы рассмотрели в этой главе, был алгоритм DBSCAN, который группирует точки, основываясь на локальных плотностях, и способен обрабатывать выбросы и идентифицировать нешаровидные фигуры.