Лесных Кирилл, БСЦ-153

# Отчет о выполнении домашнего задания №2

## Краткое описание данных

В качестве базы данных в этом задании выступает массив случайно сгенерированных чисел в диапазоне приблизительно **от -10 до 10**, с 15 знаками после запятой.

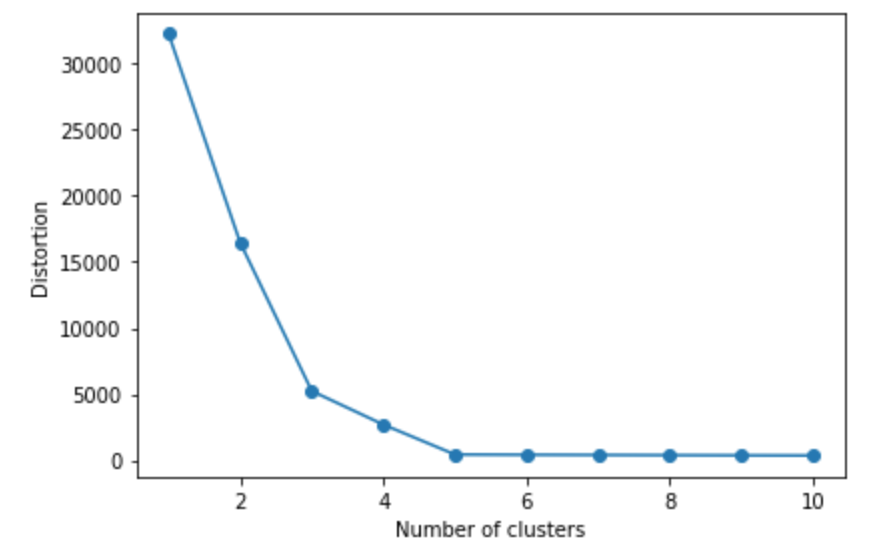
Представленный массив не несет никакого содержательного смысла, только случайный набор чисел, следовательно, никакая интерпретация данных производится не будет.

Помимо этого, в рамках данной работы не будет производиться попыток визуализировать данные, поскольку кластеризация будет происходить на основе 5 характеристике. Визуализация 5-типространственных моделей не может в должном виде быть реализована в рамках одного графика, поэтому этого действия производится не будет.

В рамках данной работы будет описано построение кластеров 3 методами: **k-means**, **иерархический** и **основанный на плотности**.

## Метод построения K-means

Прежде чем приступать к построению модели кластеризации, следует произвести разведывательный метод локтя, чтобы определить оптимальное количество кластеров в модели.



*График 1:* Метод локтя

На основе локтевого метода (*График 1*) предполагаемым оптимальным количеством кластеров является 3, 4 или 5.

Были построены 3, 4 и 5-кластерные модели. Построенные модели были проанализированы на основе показателя суммы квадратов расстояний между кластерами. Наиболее оптимальным количеством кластеров является 5, ее сумма квадратов самая наименьшая (443.87 < 2705.32 < 5242.82).

## Иерархический метод кластеризации

На основе иерархического метода кластеризации была построена следующая модель (*График 2*).



*График 2:* Дендрограмма модели иерархического анализа

На основе визуального анализа дендрограмы можно сделать вывод о том, что 5-кластерная модель наиболее оптимально и равномерно описывает массив данных.

В рамках оценки качества построенной иерархической модели был проведен анализ силуэтов (*График 3*).



*График 3:* Силуэты иерархической 5-ти кластерной модели

Силуэты представляют собой деление разности суммы расстояний между всеми точками каждого кластера на максимальное значение одной из них.

Силуэты разделены равномерно и наполнены в равной степени. На основе визуального анализа силуэтов, можно сделать вывод, что в 5-ти кластерной иерархической модели кластеры являются однородными и равноудаленными относительно друг друга.

## Метод кластеризации, основанный на плотности

Для построения модели кластеризации на основе плотности требуется установить параметры модели.

В качестве параметра «максимального расстояния между 2 выборками для объединения» (**eps)** было выбрано значение 1,6. Это является минимальным значением этого параметр, при котором были кластеризованы все наблюдения массива.

В качестве параметра «количества наблюдений в диапазоне для утверждения пиковой точки» (**min\_samples**) было выбрано значение 5. Это значение является одним из наиболее часто применяемых в построении кластерной модели на основе плотности.

В рамках построения этой модели данные разделились на 5 кластеров.

*Таблица 1*

Принадлежность всех наблюдений к кластерам  
в модели основанной плотности

