**Лабораторная работа №3**

КЛАСТЕРИЗАЦИЯ

По дисциплине «Машинное обучение»

Выполнил студент

группы 3530904/80102: Шерман М.Л.

Преподаватель: Селин И.А.

Оглавление

[Задачи 3](#_Toc71472870)

[Пункт 1 4](#_Toc71472871)

[Пункт 2 6](#_Toc71472872)

[Первый датасет 6](#_Toc71472873)

[Второй датасет 8](#_Toc71472874)

[Третий датасет 9](#_Toc71472875)

[Пункт 3 12](#_Toc71472876)

[Пункт 4 14](#_Toc71472877)

[Вывод 15](#_Toc71472878)

[Приложение 16](#_Toc71472879)

# Задачи

1. Разбейте множество объектов из набора данных pluton.csv на 3 кластера с помощью k-means. Сравните качество разбиения в зависимости от максимального числа итераций алгоритма и использования стандартизации.
2. Разбейте на кластеры множество объектов из наборов данных clustering\_1.csv, clustering\_2.csv и clustering\_3.csv с помощью k-means, DBSCAN и иерархической кластеризации. Определите оптимальное количество кластеров (где это применимо). Какой из методов сработал лучше и почему?
3. Осуществите сжатие цветовой палитры изображения (любого, на ваш выбор). Для этого выделите n кластеров из цветов всех пикселей изображения и зафиксируйте центра этих кластеров. Создайте изображение с цветами из сокращенной палитры (цвета пикселей только из центров выделенных кластеров). Покажите исходное и сжатое изображения.
4. Постройте дендрограмму для набора данных votes.csv (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год). Строки представляют 50 штатов, а столбцы - годы выборов (31). Проинтерпретируйте полученный результат.

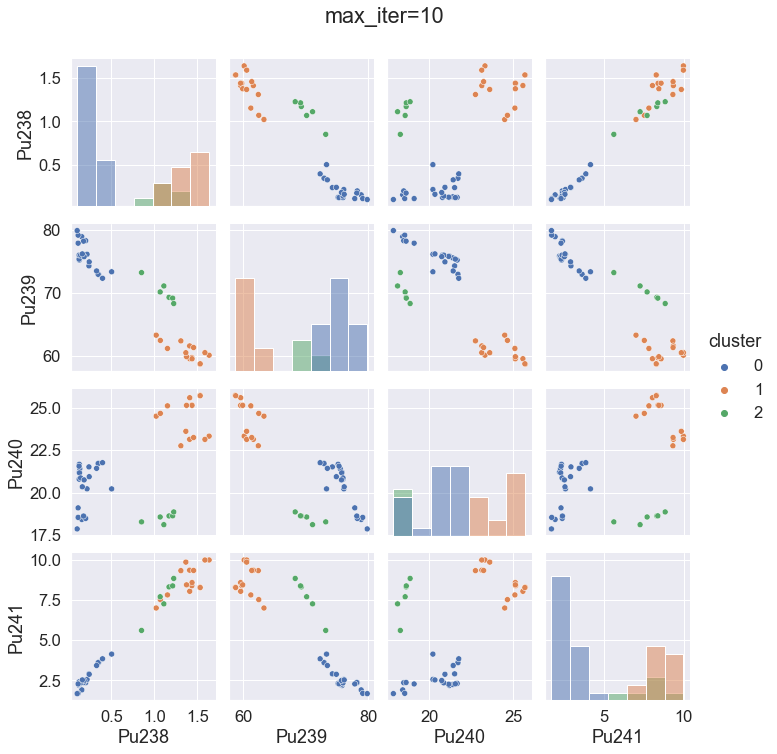
# Пункт 1

В данной лабораторной работе необходимо изучить одну из задач обучения без учителя, а именно кластеризацию. Для изучения основных алгоритмов и методов оценки качества разбиения воспользуемся библиотекой scikit-learn.

В первом пункте рассмотрим качество разбиения метода k средних, в зависимости от максимального количества итераций метода. Для оценки качества разбиения воспользуемся метрикой силуэта (silhouette\_score). Данная метрика позволяет оценить качество кластеризации, используя только саму неразмеченную выборку и результат кластеризации.

Будем рассматривать следующий набор итераций: 1, 10, 100, 1000.

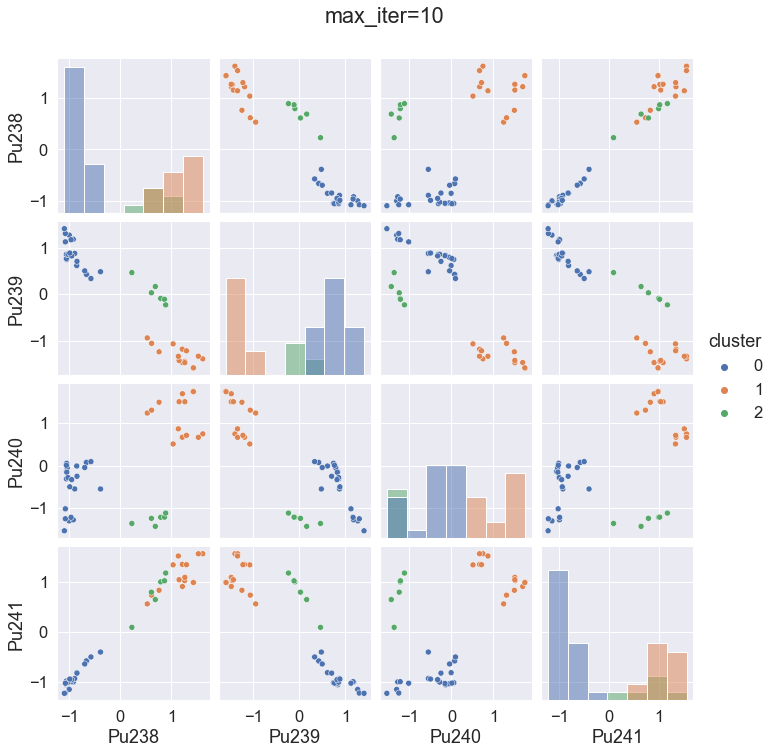
|  |  |  |
| --- | --- | --- |
| Максимальное число итераций | Потребовавшееся число итераций | Силуэт |
| 1 | 1 | 0.6782467064358735 |
| 10 | 3 | 0.6782467064358735 |
| 100 | 3 | 0.6782467064358735 |
| 1000 | 3 | 0.6782467064358735 |

Как видно из результатов, качество разбиения методом k средних для данной выборки не зависит от количества итераций. Рассмотрим наглядно, как данный метод справился с разбиением.

Взглянув на попарные распределения признаков, можно сказать, что метод k-средних достаточно хорошо справился со своей задачей, однако показатель качества не лучший. Постараемся это исправить с помощью стандартизирования выборки.

Аналогично рассмотрим набор итераций 1, 10, 100, 1000.

|  |  |  |
| --- | --- | --- |
| Максимальное число итераций | Потребовавшееся число итераций | Силуэт |
| 1 | 1 | 0.7466594999843088 |
| 10 | 2 | 0.7466594999843088 |
| 100 | 2 | 0.7466594999843088 |
| 1000 | 2 | 0.7466594999843088 |

Качество разбиения значительно улучшилось, однако всё также методу k средних не нужно большое число итераций. Рассмотрим график:

Визуально отличий не наблюдается, однако по сетке видно, что стандартизация помогла более качественно изображать найденные классы.

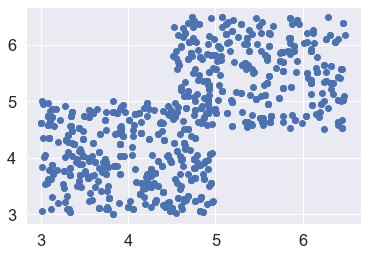
# Пункт 2

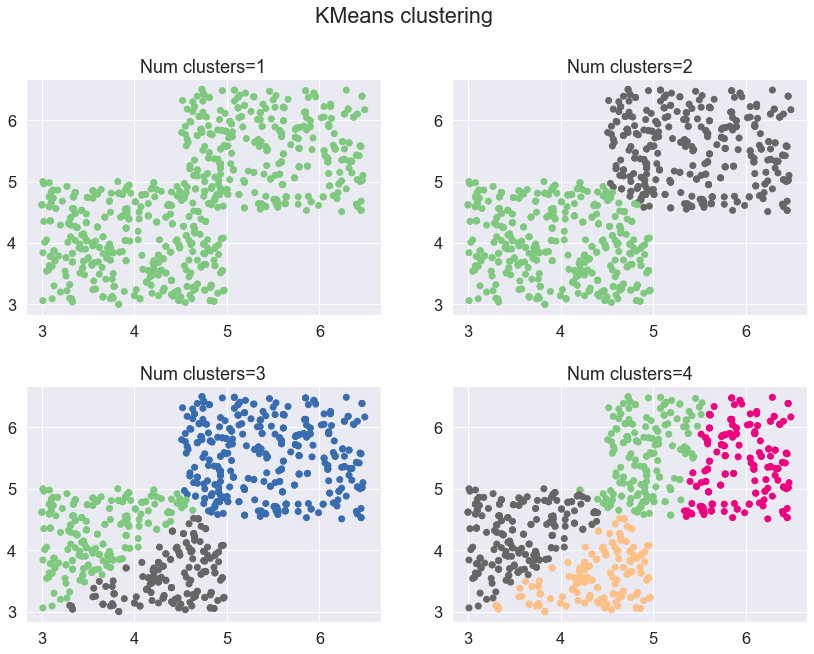
Рассмотрим на новых датасетах работу трёх методов кластеризации:

* Метод k средних
* Основанная на плотности пространственная кластеризация для приложений с шумами (DBSCAN)
* Иерархическую кластеризацию

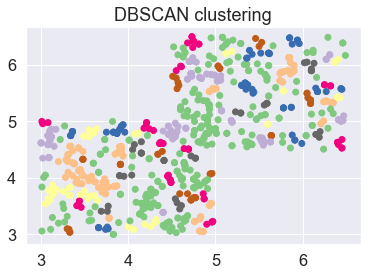
Метод k средних и иерархическая кластеризация требуют передавать количество кластеров, на которые надо разделить переданную выборку. Для более детального изучения данных методов будем рассматривать их работу с параметром n\_clusters ϵ {1, 2, 3, 4}.

## Первый датасет

Рассмотрим «расположение» объектов на плоскости:

Визуально различимы два кластера. Посмотрим, как справится с разбиением метод k средних:

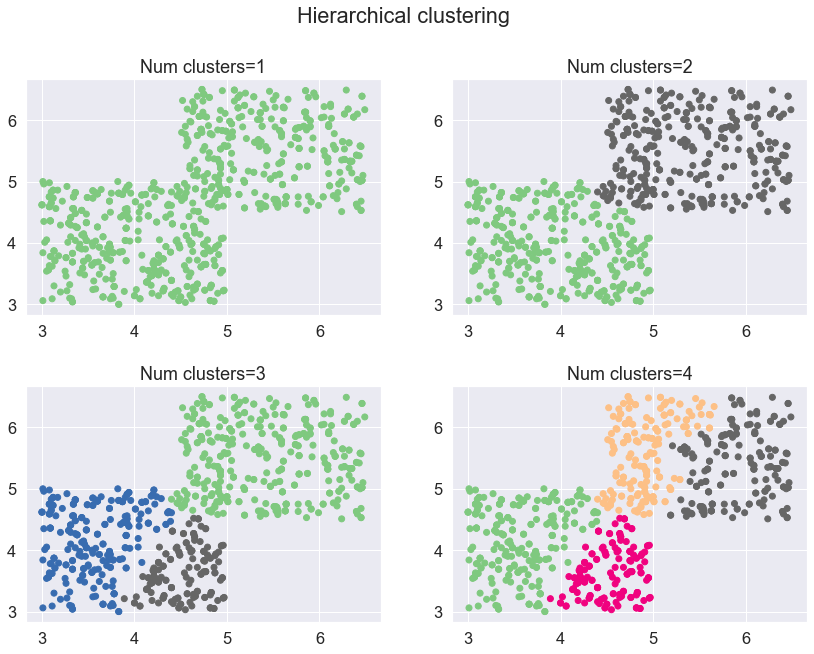
Метод k средних прекрасно справился с поставленной задачей. Конечно, при избыточном значении желаемого количества кластеров, результаты не совсем те, которые ожидались, однако и слишком странными (с визуальной точки зрения) их не назвать.

Рассмотрим результат DBSCAN:

В данном алгоритме изменялись два параметра:

* eps – максимальное расстояние между двумя выборками
* min\_samples – количество выборок (или общий вес) в окрестности точки, рассматриваемой в качестве основной

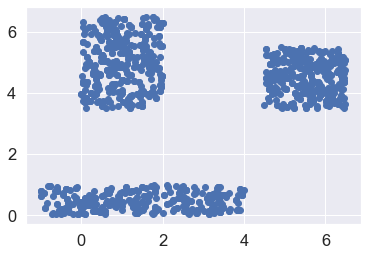
Любая комбинация данных параметров не приводила к удовлетворительному результату. Это связано с «наложением» части одного кластера на другой.

Рассмотрим результат иерархической кластеризации:

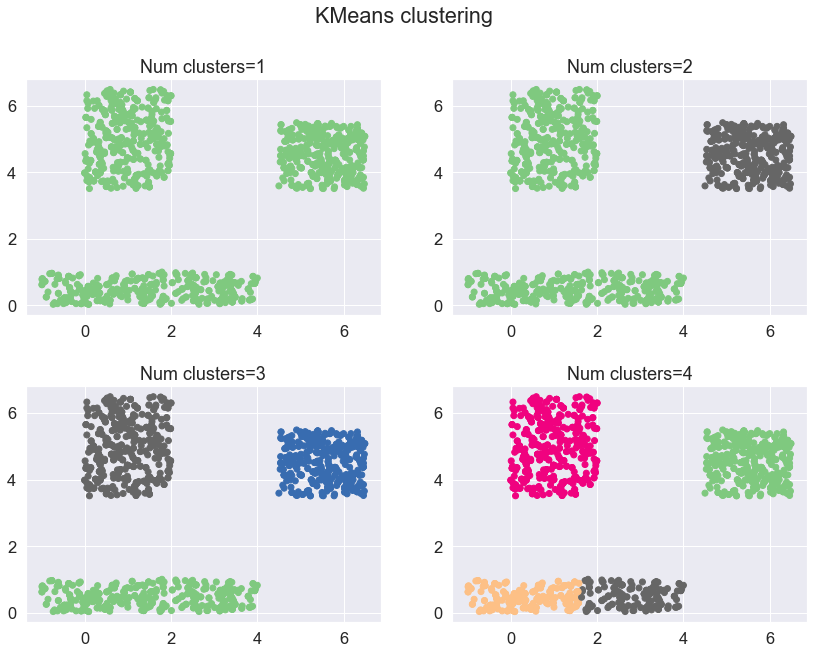
Результат практически аналогичен методу k средних.

В данном методе можно задать метрику, для вычисления связи между выборками. Оптимальным вариантом для данного набора данных стала метрика ward, которая минимизирует сумму квадратов различий во всех кластерах. Данный подход минимизирует дисперсию, и в этом смысле он похож на метод k средних, что и объясняет схожесть разбиений.

## Второй датасет

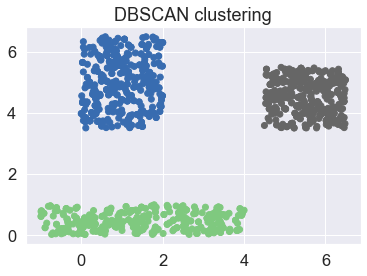
Рассмотрим «расположение» объектов на плоскости:

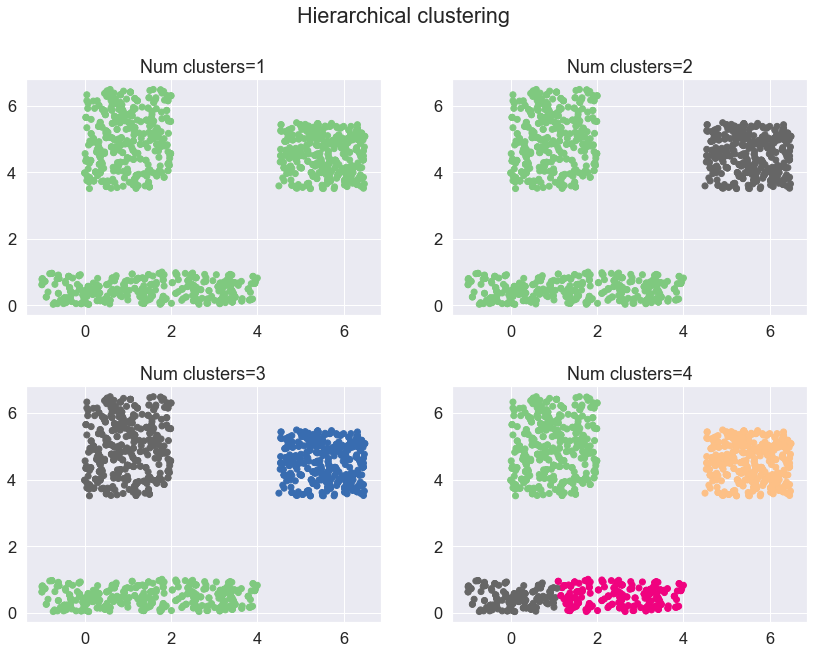
Хорошо различимы три выборки. Испытаем на данном датасете наши методы.

Метод k средних:

Прекрасный результат, однако всё те же замечания о при избыточном значении желаемого количества кластеров.

DBSCAN:

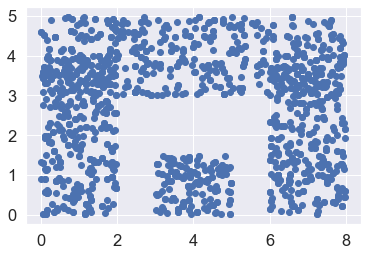
Поскольку все три кластера обособлены друг от друга, данный метод отлично справился с задачей. Используемые параметры: eps= 0.3, min\_samples=1.

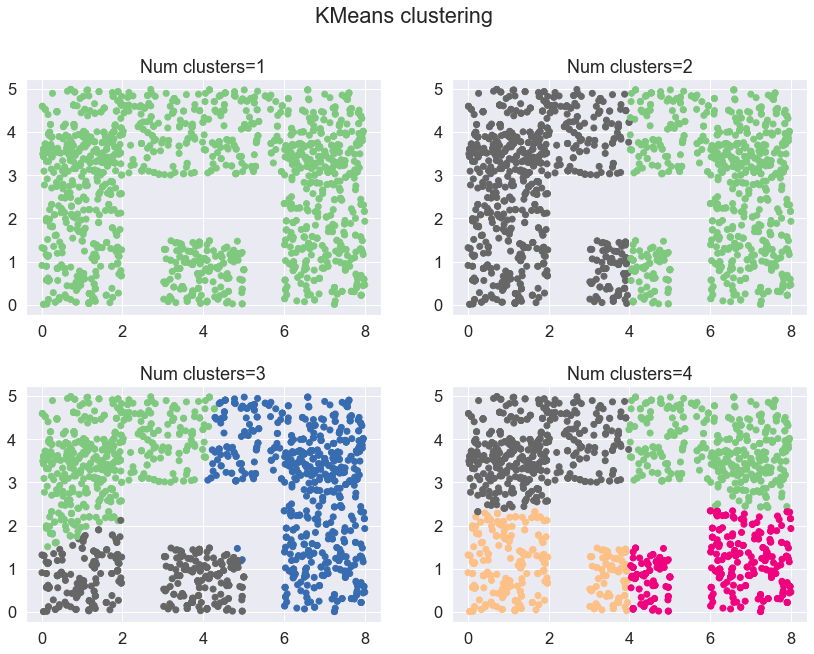
Иерархическая кластеризация:

Как и в прошлый раз использовали метрику ward, для вычисления связи между выборками. Результат схож с методом k средних и также показывает отличный результат.

## Третий датасет

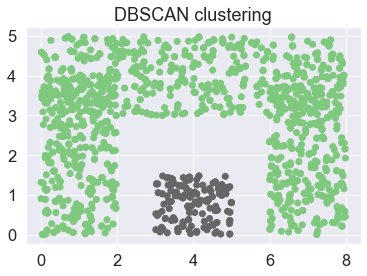
Рассмотрим «расположение» объектов на плоскости:

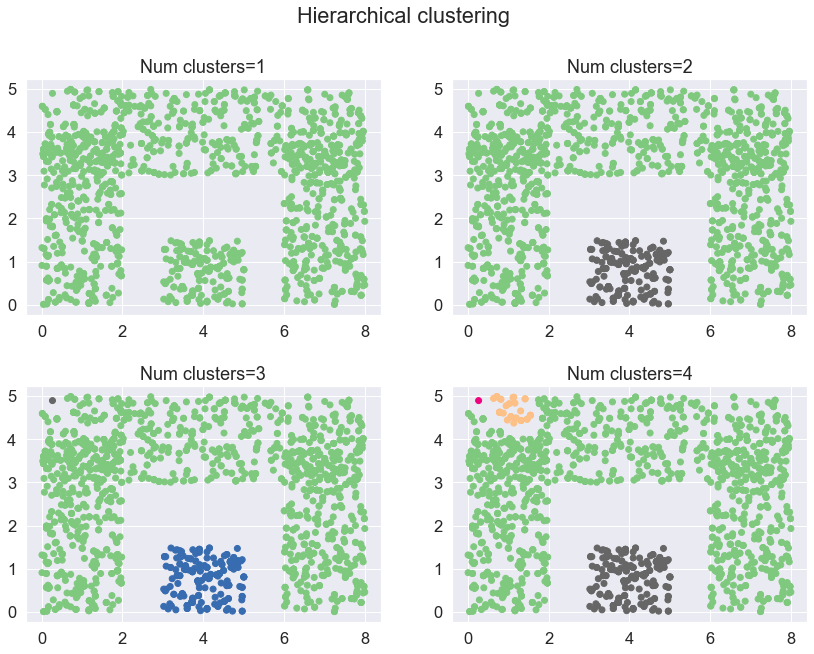
Визуально различимы два кластера, однако сложность состоит в том, что один из них частично «обрамляет» другой, из-за чего данный случай может стать серьёзным испытанием для некоторых алгоритмов.

Рассмотрим метод k средних:

К сожалению, данному методу не удалось различить два кластера. Посмотрим, как справятся остальные.

DBSCAN:

Поскольку DBSCAN ориентируется на максимальное расстояние между двумя выборками и количество выборок в окрестности точки, он смог определить два кластера при eps=0.4 и min\_samples=1.

Иерархическая кластеризация:

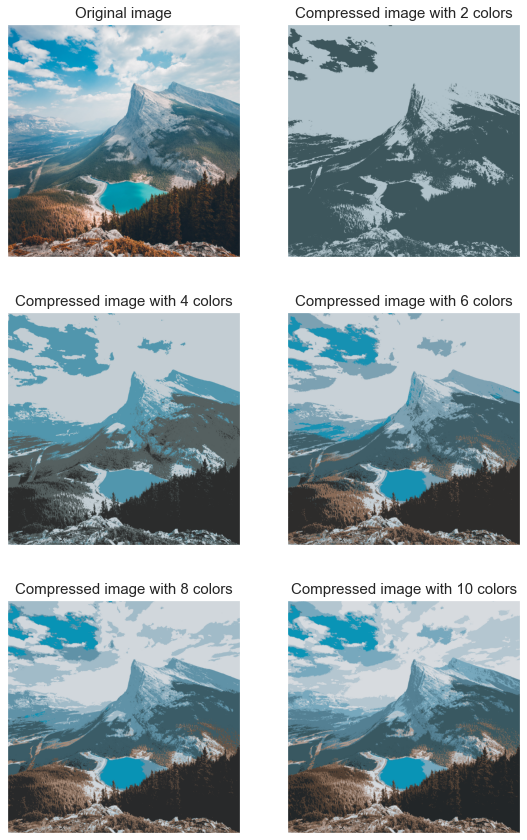
Поскольку при использовании метрики ward, поведение иерархической кластеризации схоже с методом k средних, необходимо выбрать иную метрику. Будем использовать метрику, которая минимизирует расстояние между ближайшими наблюдаемыми парами кластеров. С данной метрикой иерархическая кластеризация справляется с поставленной задачей.

В результате, были рассмотрены три метода кластеризации. Метод DBSCAN хорошо показывает себя, когда классы обособлены друг от друга и мы не знаем их точного количества. Метод k средних и метод иерархической кластеризации (с метрикой ward) показывают аналогичные результаты, однако второй метод поддерживает различные метрики, которые позволяют качественно работать с большим разнообразием распределения данных.

# Пункт 3

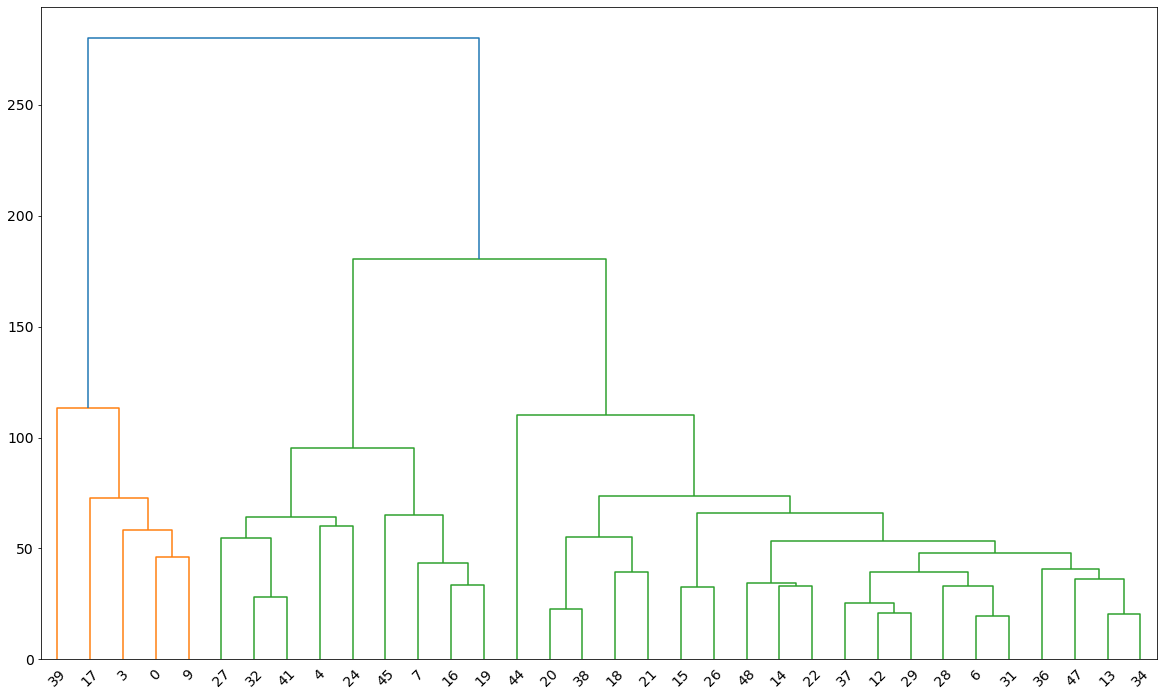
Рассмотрим применение кластеризации для выделения n цветов из всех пикселей изображения и окрасим исходное изображения цветами из сокращённой палитры. Исходное изображение:

Посмотрим, как будет изменяться изображение при 2, 4, 6, 8 и 10 цветах в палитре.



С увеличением количества цветов в палитре, изображения приобретает всё больше деталей. Так, с двумя цветами видны лишь очертания горы и долины. С четырьмя цветами можно различить озеро и небо, ландшафт становится более чётким. При шести цветах появляются оттенки коричневого, позволяющие лучше различать лес и наполнение долины. Палитры с восьмью и десятью цветами добавляют тени, придавая объём и выделяя детали картины.

# Пункт 4

Рассмотрим предоставленные данные. Больше половины штатов не имеют полных данных (скорее всего зависит от года образования штата). Лишь о 19 штатах есть полная информация. Дополним это число, заполнив пропущенные значения некоторых штатов медианой процента голосов (возьмём только те, что имеют не более 3 пропущенных значений). В результате получим информацию о 34 штата. Построим дендограмму для получившегося набора данных.

Из дендограммы видно, что методу иерархической кластеризации удалось разделить данные на два кластера, однако результат нельзя назвать точным, из-за достаточно близкого расстояния между всеми разветвлениями. Из-за этого нельзя точно определить нужное ограничение количества кластеров.

# Вывод

В результате проделанной работы удалось познакомиться с одной из задач обучения без учителя – кластеризацией. Были рассмотрены три алгоритма: k средних, DBSCAN и иерархическая кластеризация, а также методы оценки качества, такие как силуэт и наглядное изучение разбиения. Удалось изучить работу кластеризации не только на привычных наборах данных, но и на изображениях.

# Приложение

Весь код и графики можно найти в следующем репозитории:

<https://github.com/Mark-Sherman-SE/ML-Labs>