

Федеральное государственное автономное образовательное учреждение  
высшего образования

Национальный исследовательский технологический университет

«МИСиС»

Институт ИТАСУ (ИТКН)

Кафедра Инженерной кибернетики

**Лабораторная работа №3**  
**по курсу «Нейронные сети и машинное обучение»**

Выполнил:

Студент гр. МПИ-20-4-2

Малынковский О.В.

Проверил:

Курочкин И.И.

Москва 2020г.

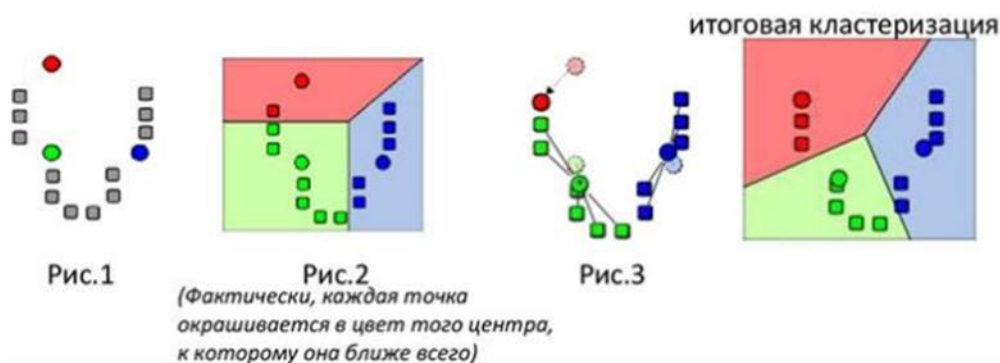
## Оглавление

Теория.....	3
Реализация.....	7
Примеры работы .....	7
Вывод.....	25
Инструкция по запуску .....	26

## Теория

Кластерный анализ предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу, были максимально «схожи», а элементы из разных групп максимально «отличались» друг от друга.

K-Means - самый популярный метод кластеризации. Метод производит разбиение исходного множества объектов на непересекающиеся кластеры и центроиды, представляющие их. Этот метод интуитивно прост: случайным образом выбираются центры кластеров, затем каждая точка из исходного множества причисляется к центру, к которому находится ближе всего. Затем центры кластеров уточняются и вместе с ними остальные элементы могут быть переназначены. Когда спустя некоторое число повторений таких уточнений центры стабилизируются, полученные центроиды и будут считаться построенной кластеризацией.



Задание этой лабораторной работы предполагает использование Евклидова расстояния и Манхэттенского. Однако k-means изначально разработан для евклидова расстояния, и, хотя можно заменить функцию расстояния в коде, тем не менее это будет плохо соотноситься с подсчетом центроидов, которые вычисляются через среднее (по сути, через расстояние евклида) точек по координатам. Такая смесь двух метрик может привести к тому, что метод не сойдется.

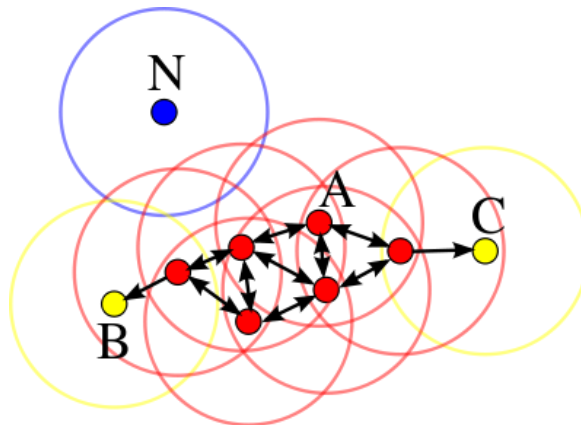
Чтобы обойти это ограничение можно использовать модифицированную версию k-means, который работает с любой функцией расстояния – k-medoids. В отличие от k-means, в k-medoids в качестве центроидов может выступать не любая точка, а только какие-то из имеющихся наблюдений. Медоиды определяются через сумму всех расстояний от одной точки до остальных в кластере: та точка, для которой такая сумма минимальна, и становится новым центром тяжести кластера на очередной итерации.

Следующий неиерархический метод, который использовался в лабораторной работе – DBSCAN (Density-based spatial clustering of applications with noise, плотностной алгоритм пространственной кластеризации с присутствием шума), как следует из названия, оперирует плотностью данных. На вход он принимает матрицу близости и два параметра — радиус  $\varepsilon$ -окрестности и  $minPts$  - количество соседей, минимальное для образования плотной области.

Точки делятся на *основные точки*, достижимые по плотности *точки* и *выбросы* следующим образом:

- Точка  $p$  является основной точкой, если как минимум  $minPts$  точек находятся на расстоянии, не превосходящем  $\varepsilon$  ( $\varepsilon$  является максимальным радиусом соседства от  $p$ ), до неё (включая саму точку  $p$ ). Говорят, что эти точки *достижимы прямо* из  $p$ .
- Точка  $q$  прямо достижима из  $p$ , если точка  $q$  находится на расстоянии, не большем  $\varepsilon$ , от точки  $p$  и  $p$  должна быть основной точкой.
- Точка  $q$  *достижима* из  $p$ , если имеется путь  $p_1, \dots, p_n$  с  $p_1 = p$  и  $p_n = q$ , где каждая точка  $p_{i+1}$  достижима прямо из  $p_i$  (все точки на пути должны быть основными, за исключением  $q$ ).
- Все точки, не достижимые из основных точек, считаются *выбросами*.

Теперь, если  $p$  является основной точкой, то она формирует *кластер* вместе со всеми точками (основными или неосновными), достижимые из этой точки. Каждый кластер содержит по меньшей мере одну основную точку. Неосновные точки могут быть частью кластера, но они формируют его «край», поскольку не могут быть использованы для достижения других точек.

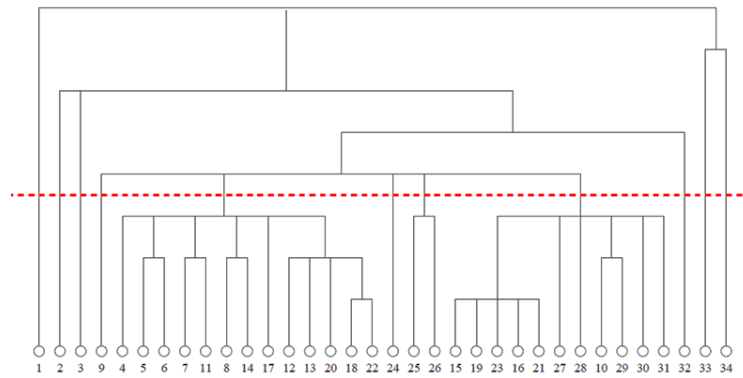


Точка A и другие красные точки являются основными точками, поскольку область с радиусом  $\varepsilon$ , окружающая эти точки, содержит по меньшей мере 4 точки (включая саму точку). Поскольку все они достижимы друг из друга, точки образуют один кластер. Точки B и C основными не являются, но достижимы из A (через другие основные точки), и также принадлежат кластеру. Точка N является точкой шума, она не является ни основной точкой, ни доступной прямо.

Рассмотрим иерархический агломеративный метод:

Иерархическая кластеризация основана на идее о том, что объекты в большей степени связаны с близлежащими объектами, чем с объектами, находящимися на отдалении.

Под термином иерархия могут иметься в виду разные вещи в разных контекстах. В случае кластеризации подразумевается вложенная структура дерева решений вида как показано на рисунке:



Верхний узел, называемый корнем, представляет весь рассматриваемый набор объектов. Каждый внутренний узел иерархии имеет несколько дочерних узлов, представляющих деление на кластеры, представленные узлами, на более мелкие кластеры. Конечные узлы соответствуют единичным объектам. Итоговая кластеризация получается путем выбора места разреза на дереве.

Агломеративный метод строит иерархию кластеров, двигаясь снизу вверх, начиная с наименьших кластеров (каждая вершина графа является отдельным кластером), обычно одиночных, и последовательно на каждом шаге объединяет те кластеры, которые находятся ближе всего друг к другу. Обычно процесс заканчивается, когда все кластеры объединяются в корневой кластер, состоящий из всех элементов исходного множества.

Чтобы решить, какие объекты / кластеры должны быть объединены, нам нужна метрика для измерения сходства между ними. Существует много способов для расчета схожести между отдельными объектами, включая евклидово и манхэттенское расстояния. Также есть разные подходы для определения расстояния между кластерами, такие как расстояние ближнего соседа (single linkage), расстояние дальнего соседа (complete linkage), групповое среднее расстояние (average linkage), расстояние между центрами, расстояние Уорда (Ward's criterion).

В данной лабораторной работе используются следующие функции расстояния:

Расстояние Евклида:

$$d(X, Y) = \sqrt{\sum_{i=1}^m (X_i - Y_i)^2}$$

Манхэттенское расстояние:

$$d(X, Y) = \sum_{i=1}^m |X_i - Y_i|$$

Для оценки качества проделанной кластеризации будем использовать следующие показатели:

Внешние:

- Adjusted Rand index

Внутренние:

- Calinski-Harabasz index
- Davies–Bouldin index
- Silhouette Coefficient
- Cluster\_cohesion

Формулы их расчёта представлены в статье по ссылке (или можно увидеть их в коде):

[https://neerc.ifmo.ru/wiki/index.php?title=Оценка\\_качества\\_в\\_задаче\\_кластеризации](https://neerc.ifmo.ru/wiki/index.php?title=Оценка_качества_в_задаче_кластеризации)

## Реализация

На языке Python с использованием библиотек Sklearn, Sklearn-extra реализуем рассматриваемые методы кластеризации. Также в sklearn есть функции оценки качества, за исключением cluster cohesion, который был реализован без библиотек.

K-means – можно указать число кластеров (n\_clusters)

```
cluster.KMeans(n_clusters=3).fit(X)
```

K-means – можно указать число кластеров (n\_clusters), метрику в данной работе не меняем (манхэттенская )

```
KMedoids(n_clusters=3, metric='manhattan').fit(X)
```

DBSCAN– можно радиус окрестности eps, минимальное число соседей для получения кластера min\_samples, metric – 'euclidean' \ 'manhattan'

```
DBSCAN(eps=0.35, min_samples=57, metric='euclidean').fit(X)
```

Аггломеративная иерархическая кластеризация – можно указать affinity – 'euclidean' \ 'manhattan', linkage - расстояние ближнего соседа (single linkage), расстояние дальнего соседа (complete linkage), групповое среднее расстояние (average linkage), расстояние между центрами, расстояние Уорда (Ward's criterion), и число кластеров n\_clusters.

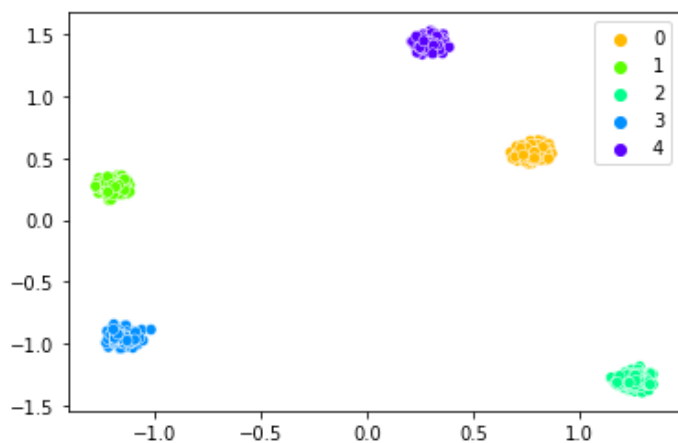
```
AgglomerativeClustering(affinity='euclidean', linkage='ward', n_clusters=3)  
.fit(X)
```

Также можно построить дерево решений:

```
model = AgglomerativeClustering(distance_threshold=0, n_clusters=None, affinity='manhattan', linkage='complete')  
model = model.fit(X)  
plot_dendrogram(model, truncate_mode='level', p=3)
```

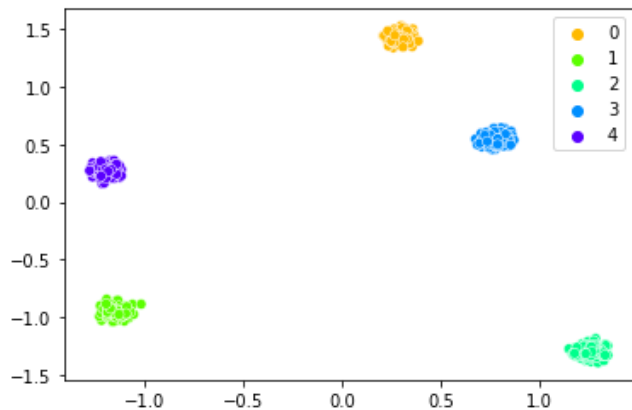
## Примеры работы

Датасет 1 - линейно разделимые множества (с расстоянием между группами в  $10^3$  раз больше, чем диаметр группы) – 5 кластеров, 1000 точек



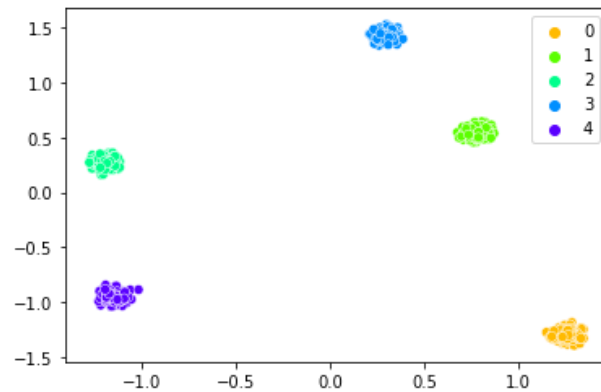
K-means:

K-means (евклидово расстояние)



Calinski-Harabasz index: 193636.11747  
 Davies-Bouldin index: 0.07375  
 Silhouette Coefficient: 0.94747  
 Cluster\_cohesion: 2.56596  
 Adjusted Rand Index: 1.0

K-medoids (манхэттенское расстояние)

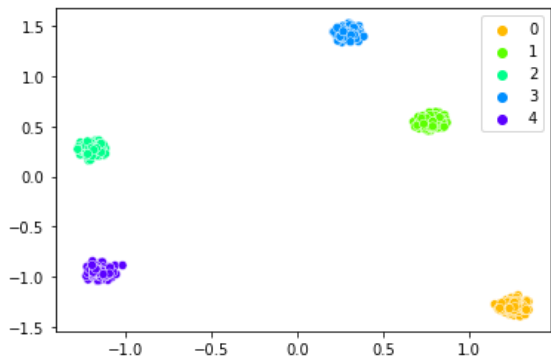


Calinski-Harabasz index: 193636.11747  
 Davies-Bouldin index: 0.07375  
 Silhouette Coefficient: 0.94413  
 Cluster\_cohesion: 2.56596  
 Adjusted Rand Index: 1.0

DBSCAN:

DBSCAN (евклидово расстояние)

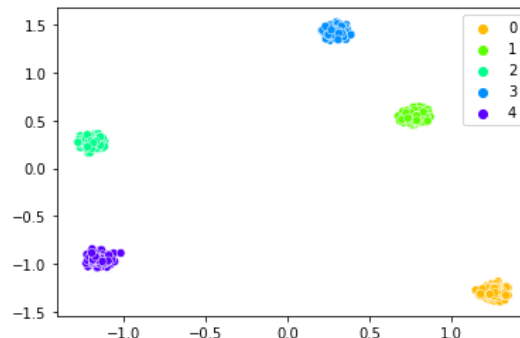
eps=0.5, min\_samples=10



Calinski-Harabasz index: 193636.11747  
 Davies-Bouldin index: 0.07375  
 Silhouette Coefficient: 0.94747  
 Cluster\_cohesion: 2.56596  
 Adjusted Rand Index: 1.0

DBSCAN (манхэттенское расстояние)

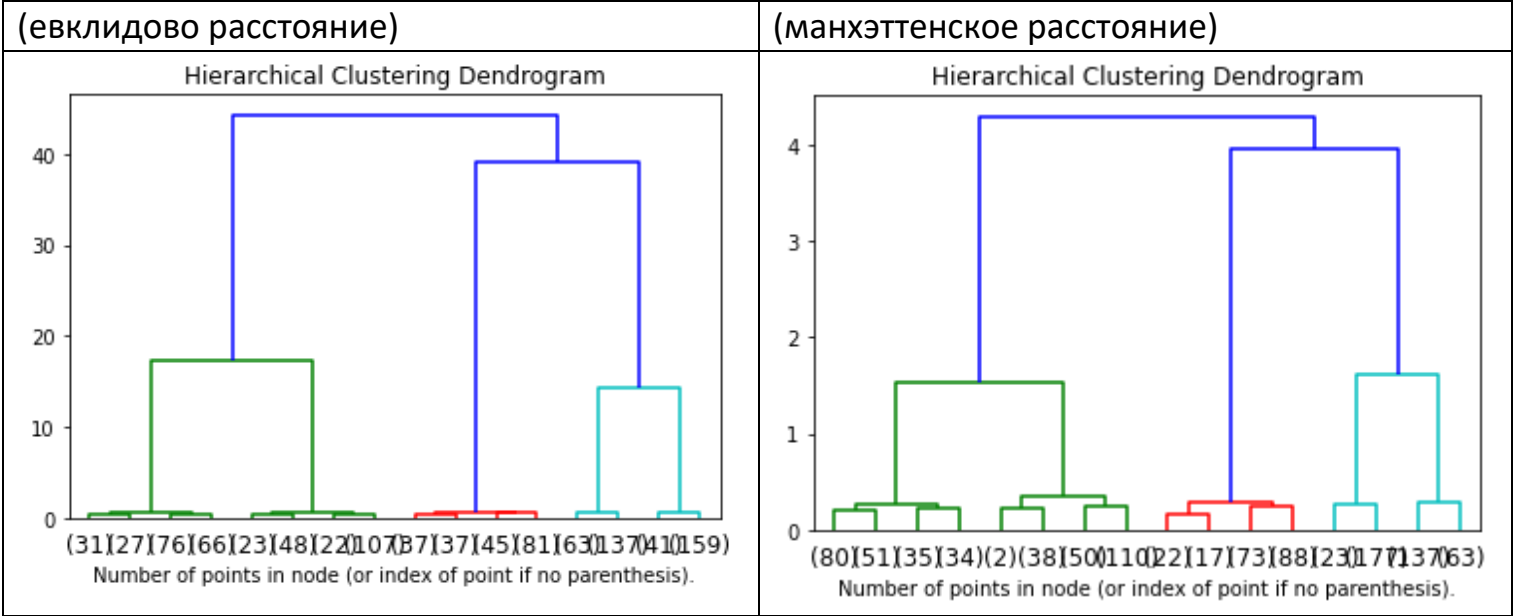
eps=0.5, min\_samples=10



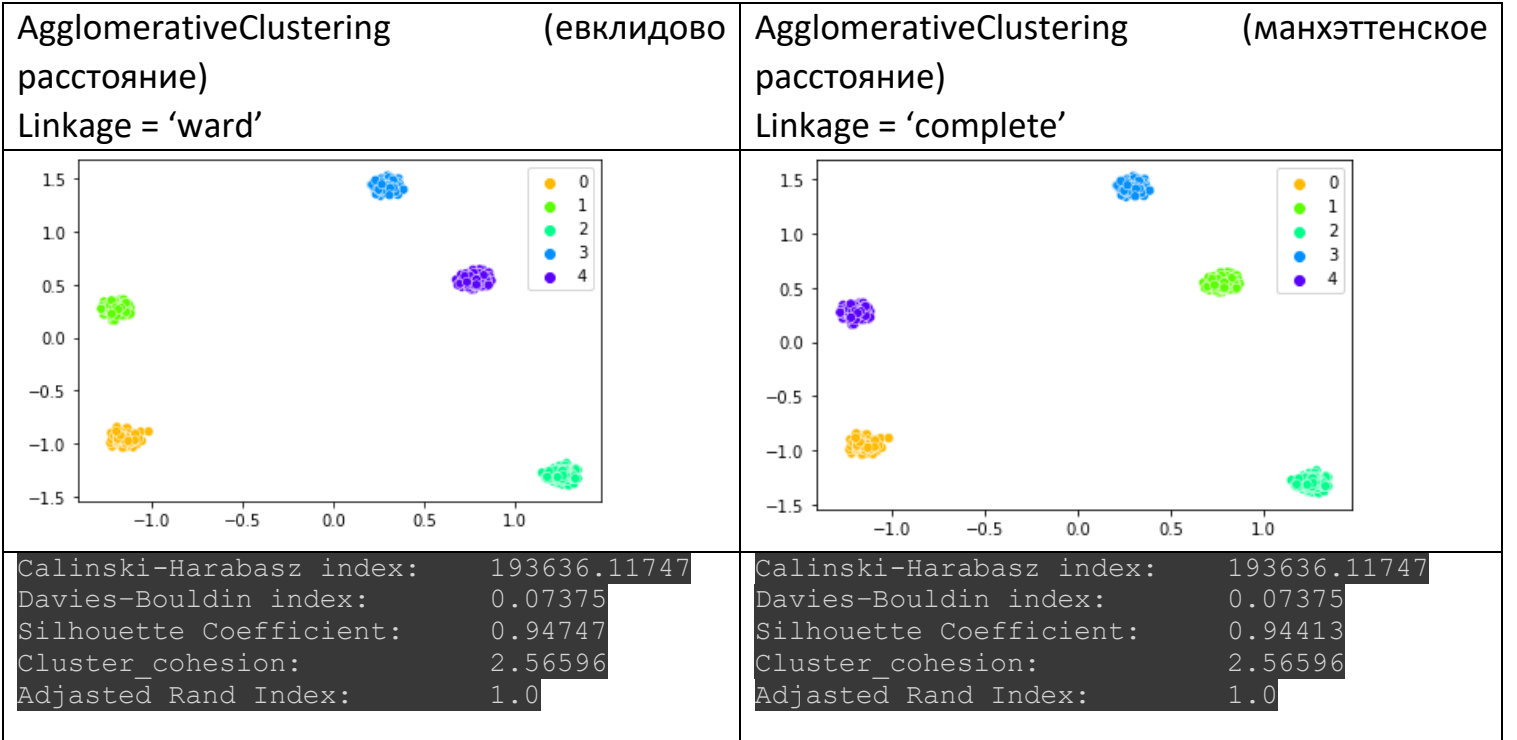
Calinski-Harabasz index: 193636.11747  
 Davies-Bouldin index: 0.07375  
 Silhouette Coefficient: 0.94413  
 Cluster\_cohesion: 2.56596  
 Adjusted Rand Index: 1.0



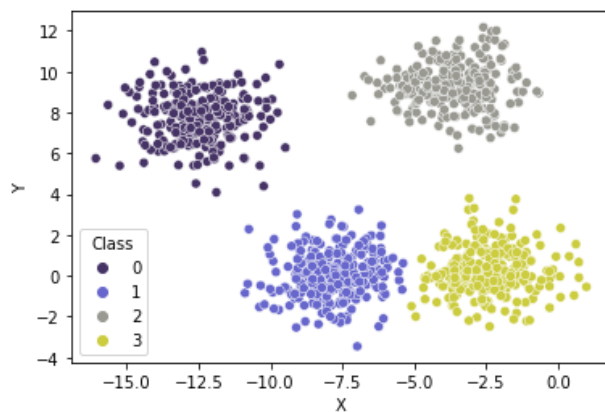
Дендрограмма:



Аггломеративная кластеризация

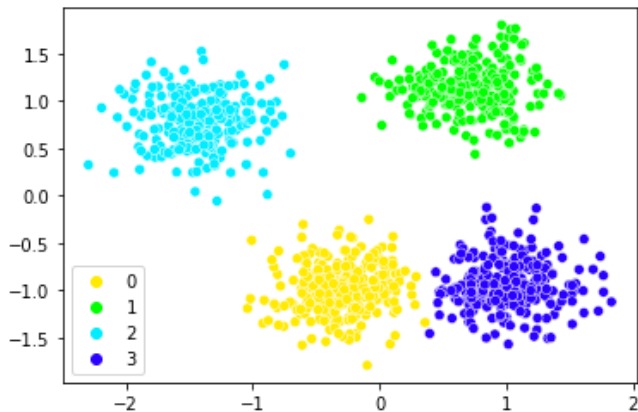


Датасет 2 - линейно разделимые множества (группы расположены близко или касаются друг друга) – 4 кластера, 1000 точек



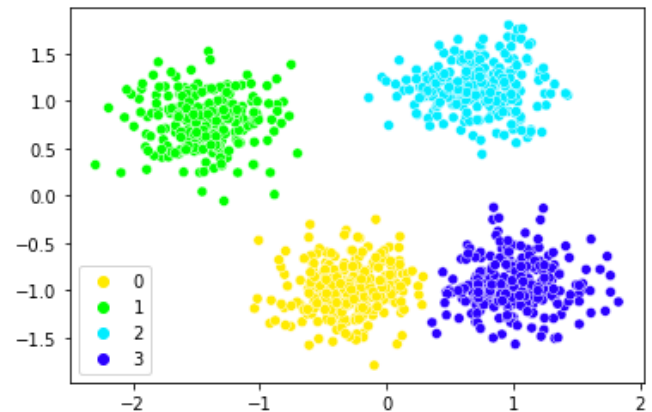
K-means:

K-means (евклидово расстояние)



Calinski-Harabasz index: 4195.79134  
 Davies-Bouldin index: 0.41994  
 Silhouette Coefficient: 0.69482  
 Cluster cohesion: 146.64987  
 Adjusted Rand Index: 0.99733

K-medoids (манхэттенское расстояние)

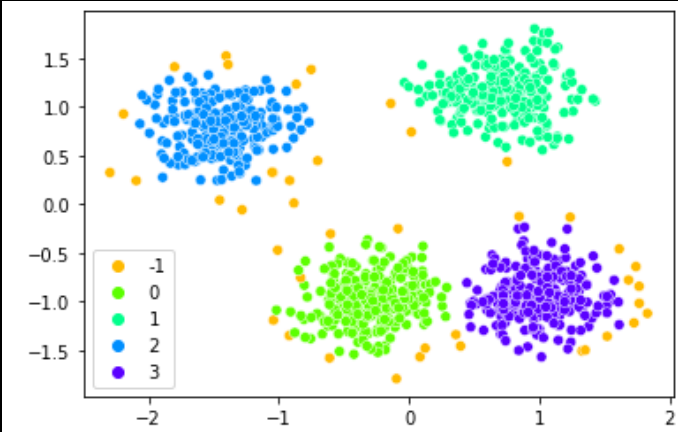


Calinski-Harabasz index: 4194.93062  
 Davies-Bouldin index: 0.41999  
 Silhouette Coefficient: 0.67309  
 Cluster cohesion: 146.67775  
 Adjusted Rand Index: 1.0

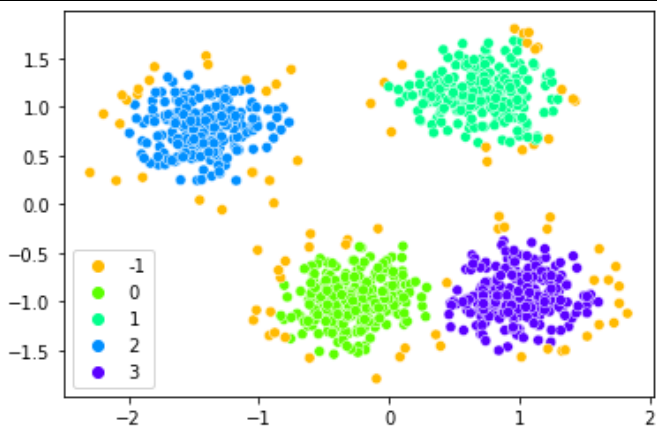
DBSCAN: (желтые точки метод определяет как выбросы)

DBSCAN (евклидово расстояние)  
 eps=0.2, min\_samples=10

DBSCAN (манхэттенское расстояние)  
 eps=0.27, min\_samples=25

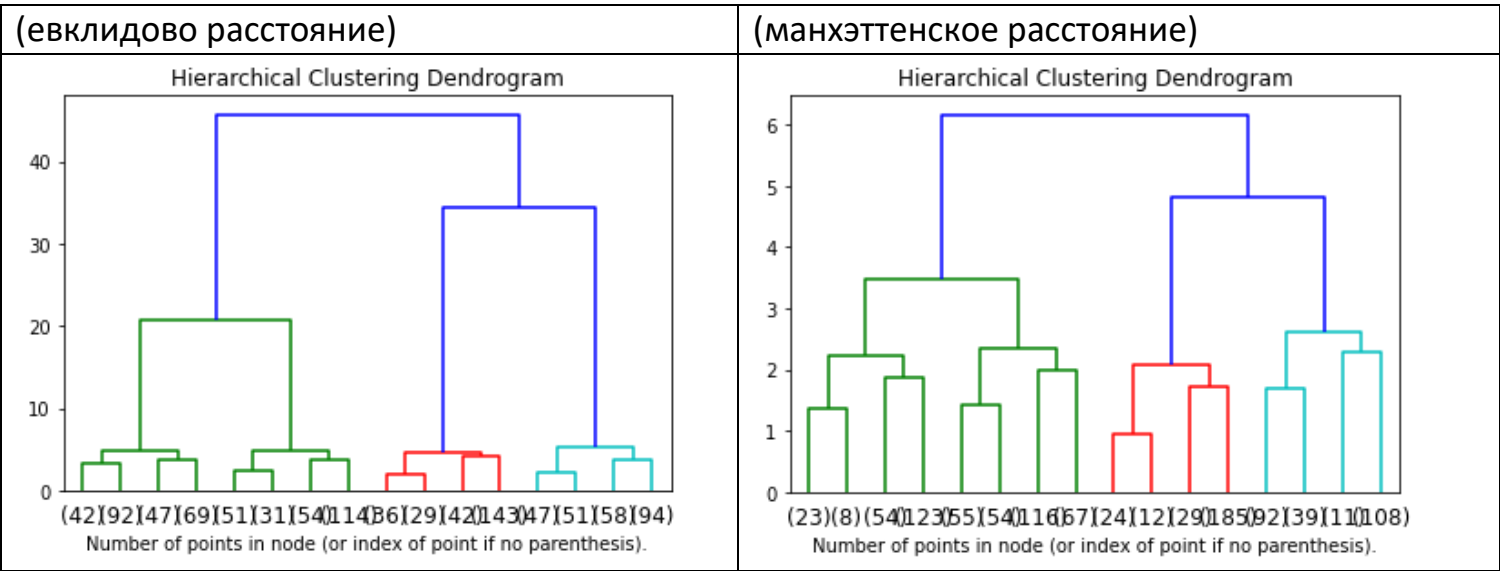


Calinski-Harabasz index: 1944.65524  
Davies-Bouldin index: 1.74070  
Silhouette Coefficient: 0.65576  
Cluster cohesion: 121.91237  
Adjusted Rand Index: 0.9419

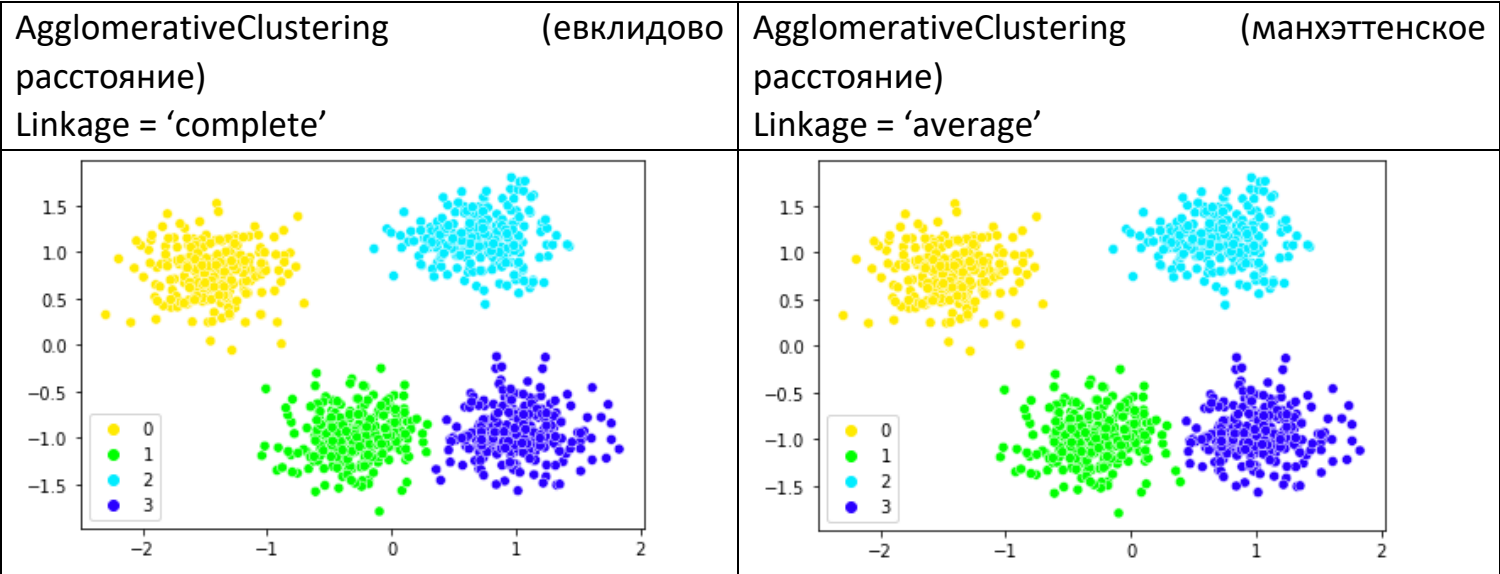


Calinski-Harabasz index: 1274.98767  
Davies-Bouldin index: 1.50287  
Silhouette Coefficient: 0.60710  
Cluster cohesion: 104.24570  
Adjusted Rand Index: 0.8792528

Дендрограмма:



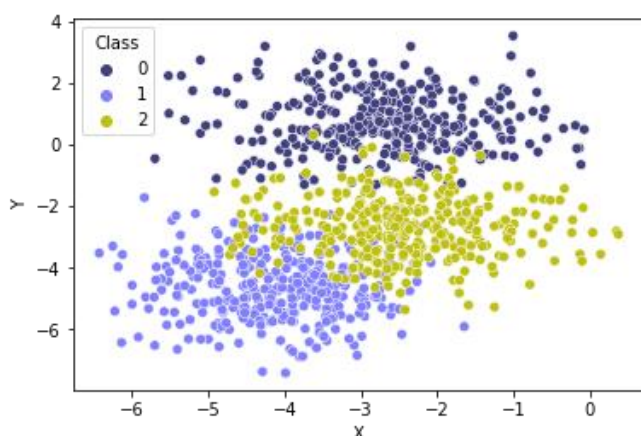
Аггломеративная кластеризация



Calinski-Harabasz index:	4194.93062
Davies-Bouldin index:	0.41999
Silhouette Coefficient:	0.69478
Cluster_cohesion:	146.67775
Adjusted Rand Index:	1.0

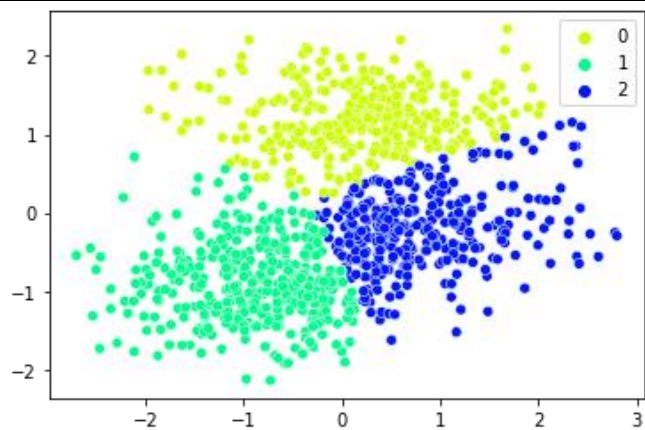
Calinski-Harabasz index:	4194.35674
Davies-Bouldin index:	0.42002
Silhouette Coefficient:	0.67301
Cluster_cohesion:	146.69635
Adjusted Rand Index:	0.994672

Датасет 3 - линейно неразделимое множество (средняя площадь пересечения классов 10-20%)– 3 кластера, 1000 точек



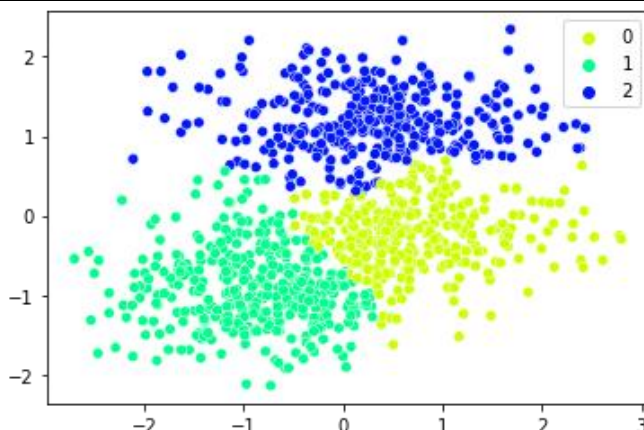
K-means:

K-means (евклидово расстояние)



Calinski-Harabasz index:	970.38623
Davies-Bouldin index:	0.87644
Silhouette Coefficient:	0.41833
Cluster_cohesion:	678.74555
Adjusted Rand Index:	0.660793

K-medoids (манхэттенское расстояние)

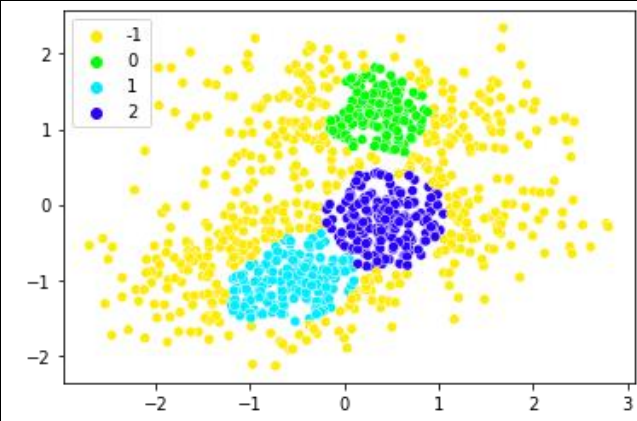


Calinski-Harabasz index:	938.35585
Davies-Bouldin index:	0.89760
Silhouette Coefficient:	0.41680
Cluster_cohesion:	693.87615
Adjusted Rand Index:	0.735976

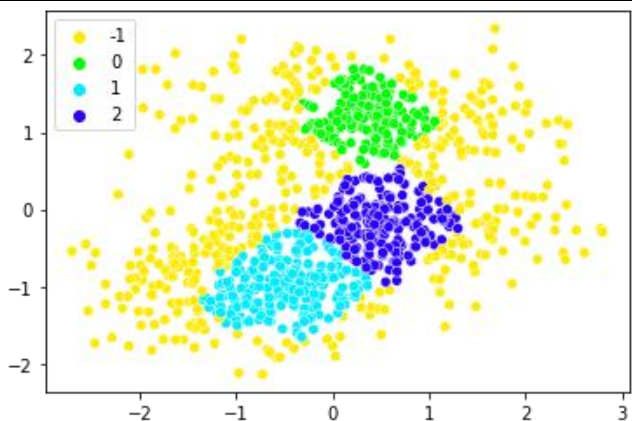
DBSCAN: (желтые точки метод определяет как выбросы)

DBSCAN (евклидово расстояние)  
eps=0.35, min\_samples=57

DBSCAN (манхэттенское расстояние)  
eps=0.5, min\_samples=70



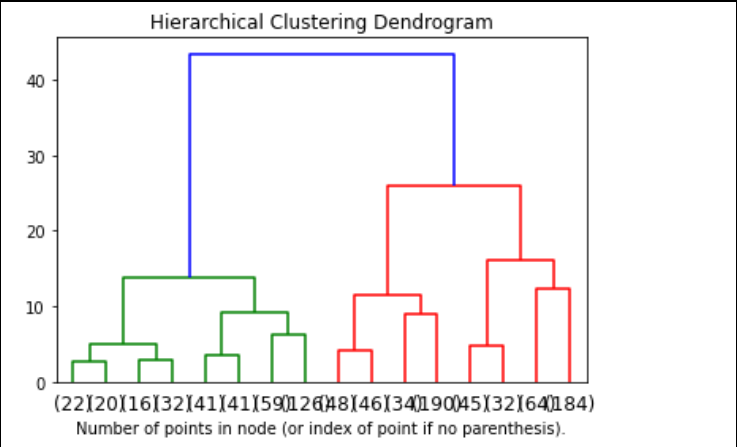
Calinski-Harabasz index: 1365.13836  
Davies-Bouldin index: 0.58730  
Silhouette Coefficient: 0.57117  
Cluster\_cohesion: 84.46853  
Adjusted Rand Index: 0.792771



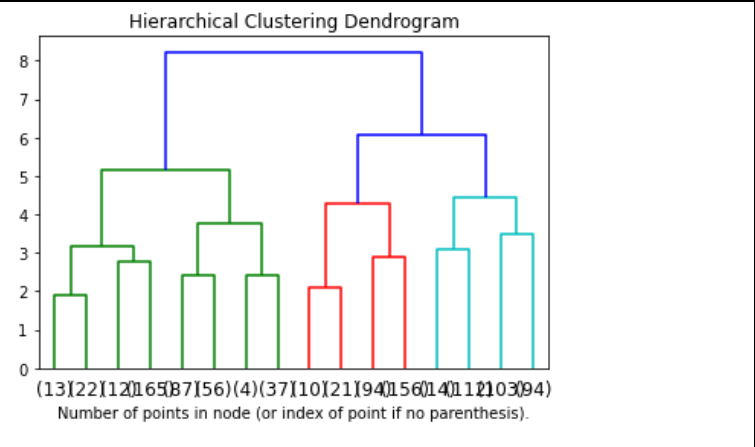
Calinski-Harabasz index: 1226.46158  
Davies-Bouldin index: 0.65764  
Silhouette Coefficient: 0.52327  
Cluster\_cohesion: 124.96802  
Adjusted Rand Index: 0.749987

Дендрограмма:

(евклидово расстояние)

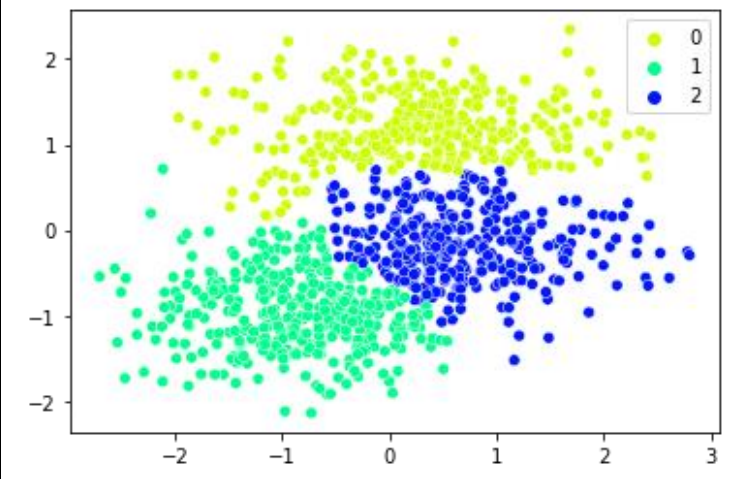


(манхэттенское расстояние)

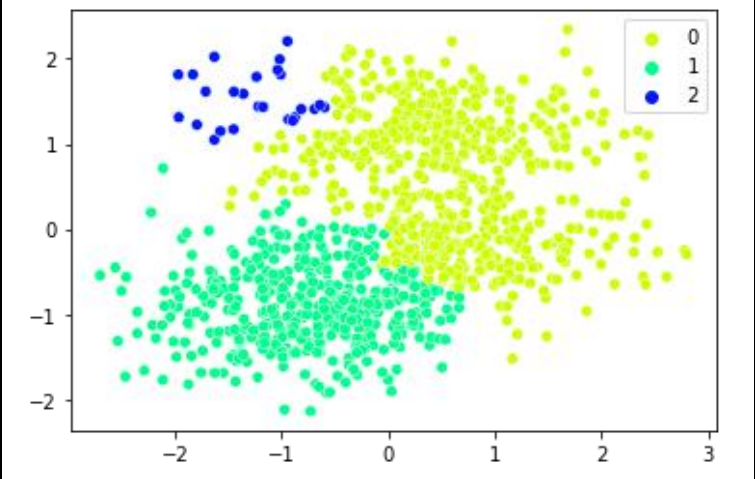


Аггломеративная кластеризация

AgglomerativeClustering (евклидово расстояние)  
Linkage = 'ward'



AgglomerativeClustering (манхэттенское расстояние)  
Linkage = 'average'

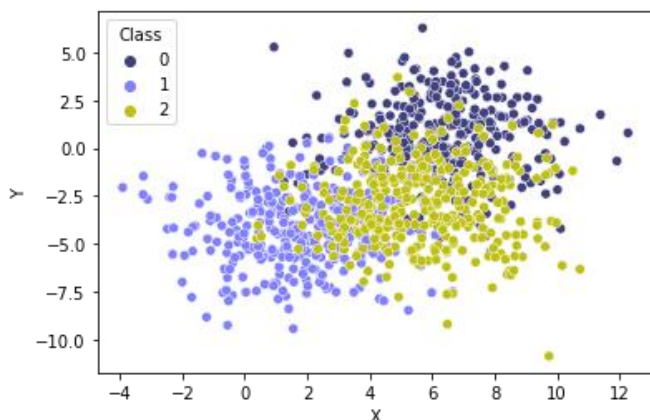




Calinski-Harabasz index:	890.60899
Davies-Bouldin index:	0.94286
Silhouette Coefficient:	0.39275
Cluster_cohesion:	717.72626
Adjusted Rand Index:	0.7345285

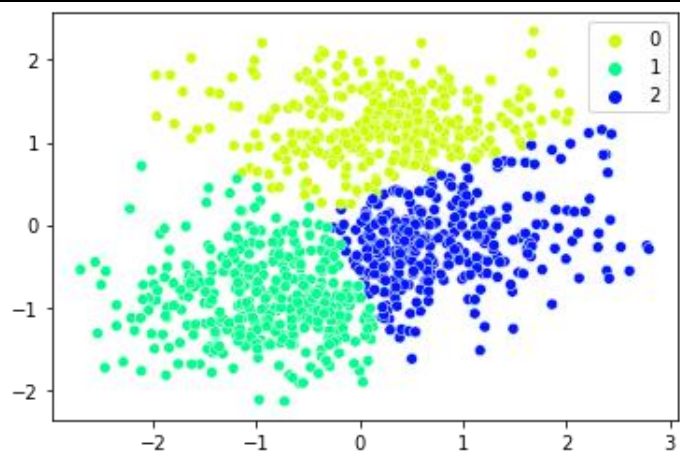
Calinski-Harabasz index:	573.82687
Davies-Bouldin index:	0.79904
Silhouette Coefficient:	0.33608
Cluster_cohesion:	929.75382
Adjusted Rand Index:	0.4338765

Датасет 4 - линейно неразделимое множество (средняя площадь пересечения классов 50-70%)– 3 кластера, 1000 точек



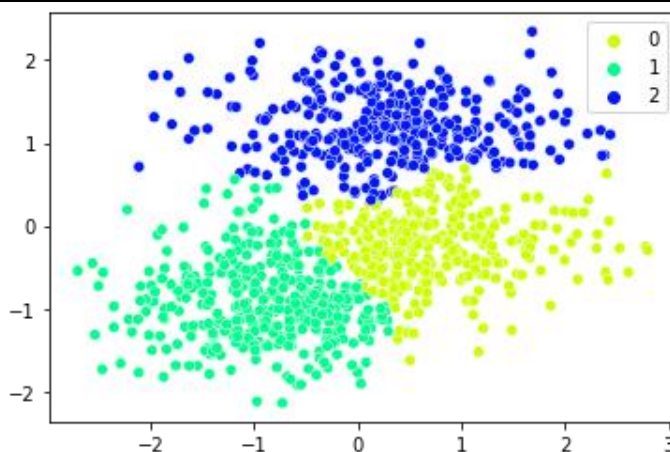
K-means:

K-means (евклидово расстояние)



Calinski-Harabasz index:	970.38623
Davies-Bouldin index:	0.87644
Silhouette Coefficient:	0.41833
Cluster_cohesion:	678.74555
Adjusted Rand Index:	0.660793

K-medoids (манхэттенское расстояние)

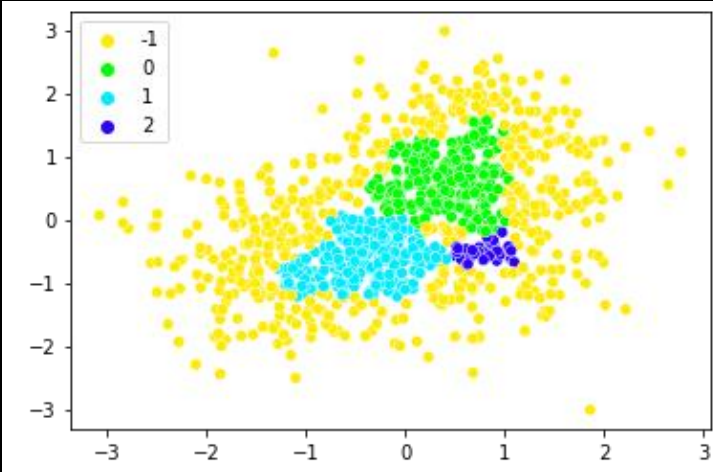


Calinski-Harabasz index:	938.35585
Davies-Bouldin index:	0.89760
Silhouette Coefficient:	0.41680
Cluster_cohesion:	693.87615
Adjusted Rand Index:	0.735976

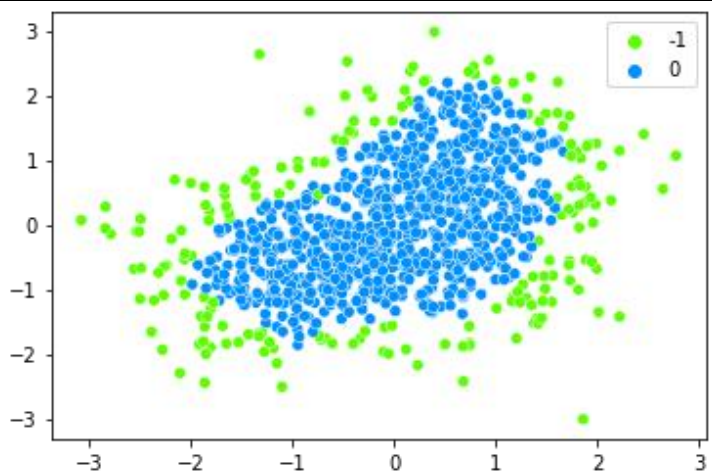
DBSCAN: (желтые точки слева (зеленые справа) метод определяет как выбросы)

DBSCAN (евклидово расстояние)  
eps=0.21, min\_samples=18

DBSCAN (манхэттенское расстояние)  
eps=0.3, min\_samples=12

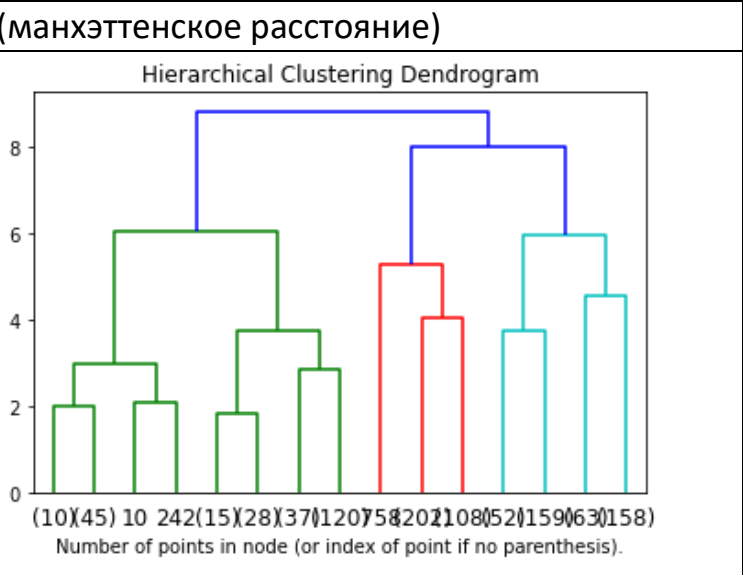
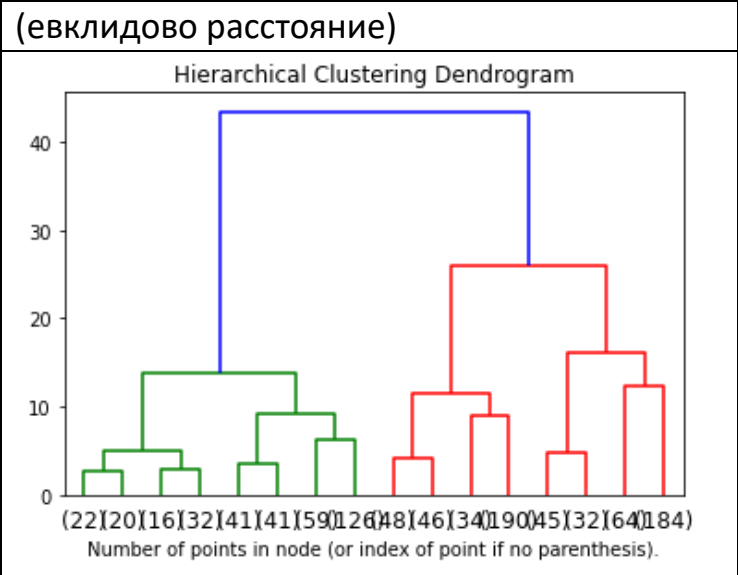


Calinski-Harabasz index: 533.83432  
Davies-Bouldin index: 0.62730  
Silhouette Coefficient: 0.43302  
Cluster cohesion: 125.94660  
Adjusted Rand Index: 0.2587415



Calinski-Harabasz index: 2.39618  
Davies-Bouldin index: 17.89319  
Silhouette Coefficient: 0.25027  
Cluster cohesion: 1141.27070  
Adjusted Rand Index: 0.00300139

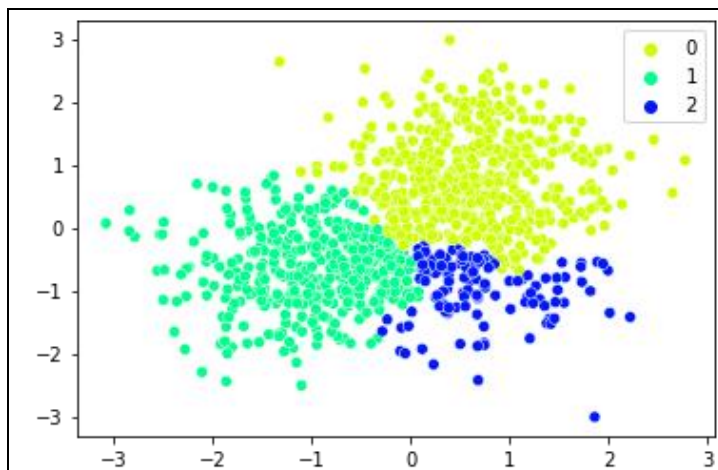
Дендрограмма:



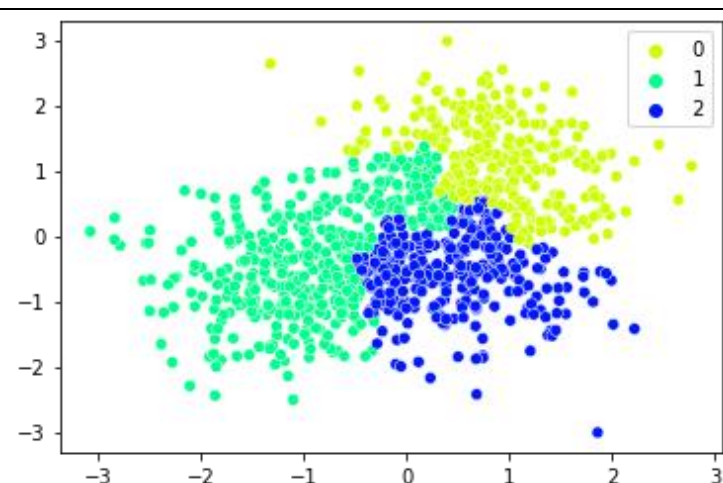
Аггломеративная кластеризация

AgglomerativeClustering (евклидово расстояние)  
Linkage = 'complete'

AgglomerativeClustering (манхэттенское расстояние)  
Linkage = 'complete'

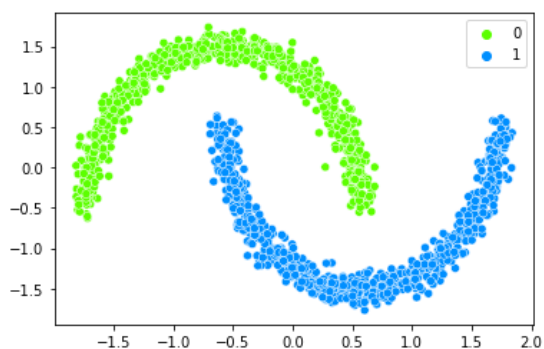


Calinski-Harabasz index: 748.03621  
 Davies-Bouldin index: 0.84534  
 Silhouette Coefficient: 0.35191  
 Cluster cohesion: 799.81632  
 Adjusted Rand Index: 0.3859005



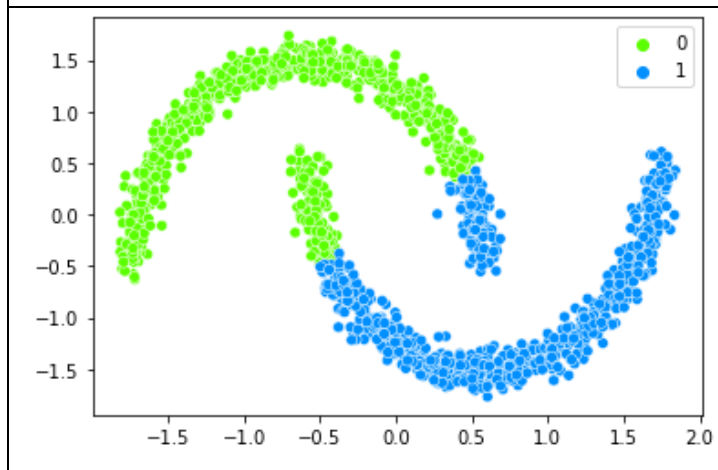
Calinski-Harabasz index: 583.14001  
 Davies-Bouldin index: 1.13538  
 Silhouette Coefficient: 0.26458  
 Cluster cohesion: 921.74844  
 Adjusted Rand Index: 0.2574876

Датасет 5 - линейно неразделимое множество без пересечений– 2 кластера, 1500 точек

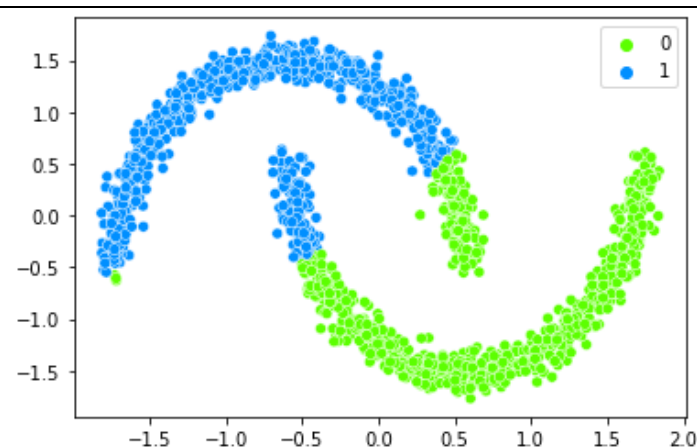


K-means:

K-means (евклидово расстояние)



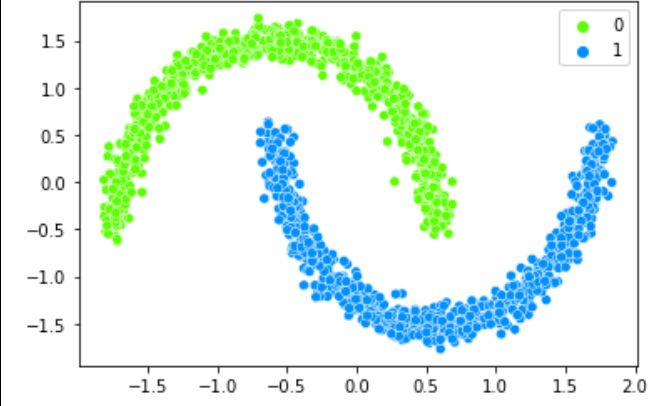
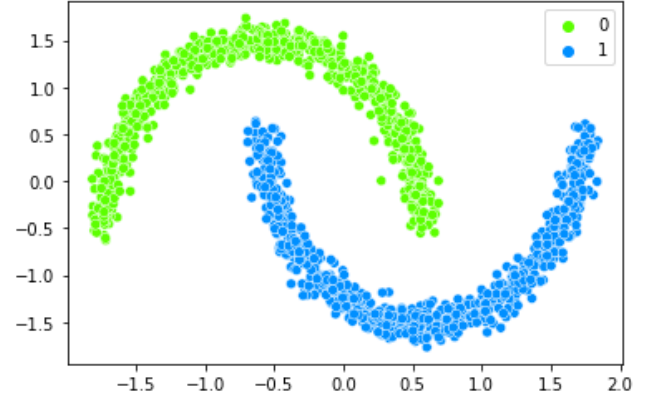
K-medoids (манхэттенское расстояние)



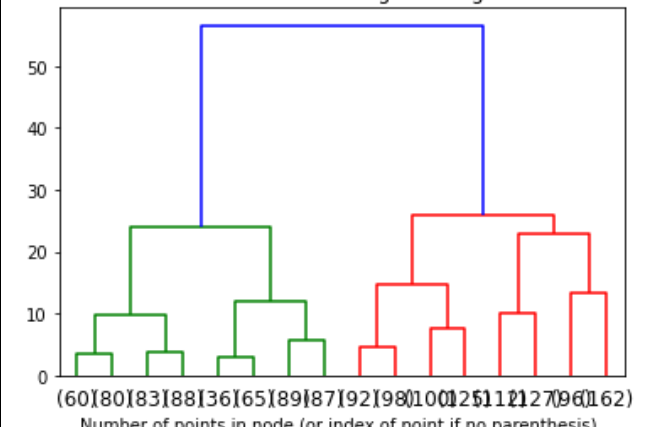
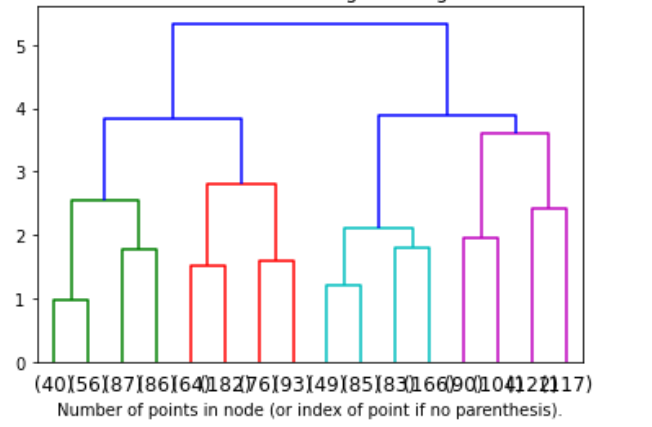


Calinski-Harabasz index: 2092.27173	Calinski-Harabasz index: 2057.26868
Davies-Bouldin index: 0.81005	Davies-Bouldin index: 0.81510
Silhouette Coefficient: 0.49709	Silhouette Coefficient: 0.49080
Cluster cohesion: 1251.71584	Cluster cohesion: 1264.03949
Adjusted Rand Index: 0.49527933	Adjusted Rand Index: 0.46204132

DBSCAN: (желтые точки слева (зеленые справа) метод определяет как выбросы)

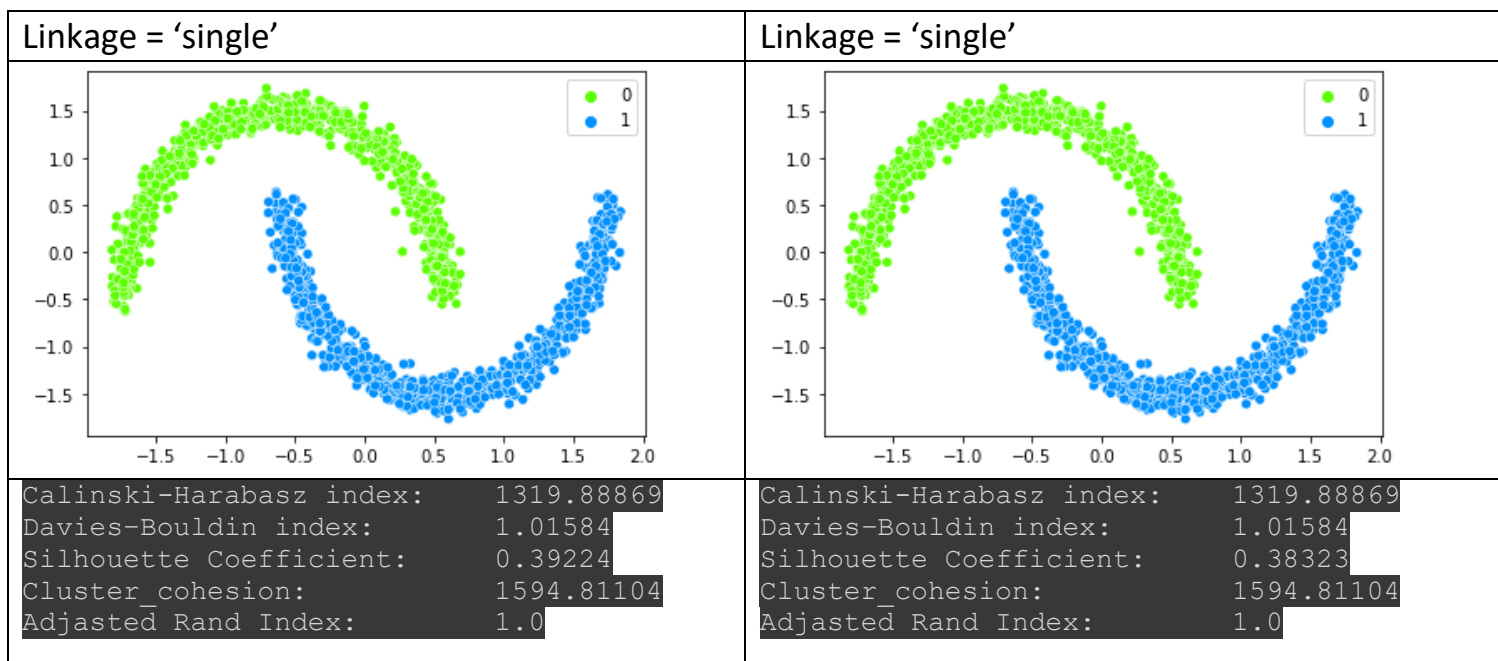
DBSCAN (евклидово расстояние) eps=0.35, min_samples=57	DBSCAN (манхэттенское расстояние) eps=0.5, min_samples=70
	
Calinski-Harabasz index: 1319.88869 Davies-Bouldin index: 1.01584 Silhouette Coefficient: 0.39224 Cluster cohesion: 1594.81104 Adjusted Rand Index: 1.0	Calinski-Harabasz index: 1319.88869 Davies-Bouldin index: 1.01584 Silhouette Coefficient: 0.38323 Cluster cohesion: 1594.81104 Adjusted Rand Index: 1.0

Дендрограмма:

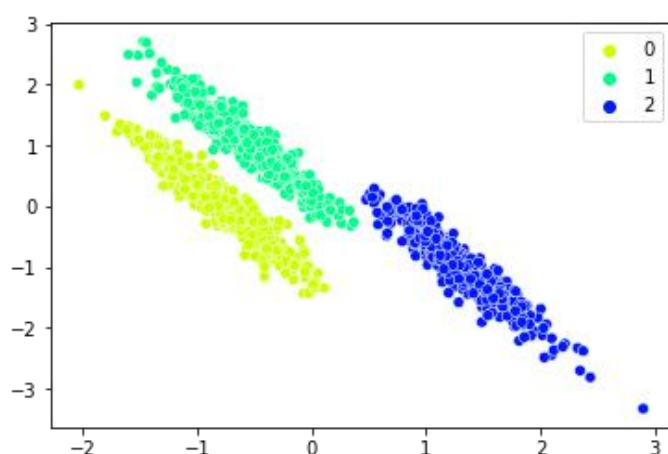
(евклидово расстояние)	(манхэттенское расстояние)
<p>Hierarchical Clustering Dendrogram</p>  <p>Number of points in node (or index of point if no parenthesis).</p>	<p>Hierarchical Clustering Dendrogram</p>  <p>Number of points in node (or index of point if no parenthesis).</p>

Агломеративная кластеризация

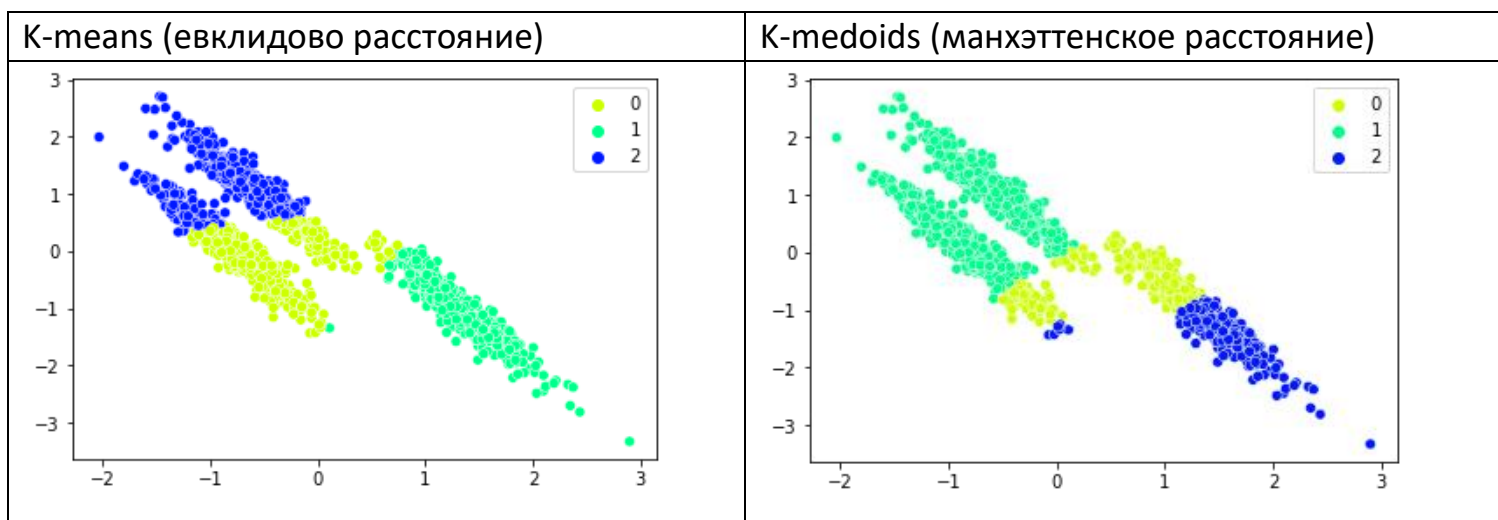
AgglomerativeClustering (евклидово расстояние)	AgglomerativeClustering (манхэттенское расстояние)
--	--



Датасет 6 - линейно разделимое множество без пересечений, но не сгущение точек не сферической формы – 3 кластера, 1500 точек

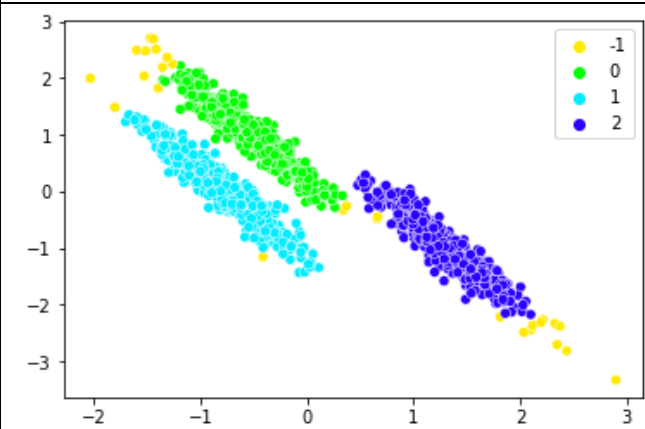
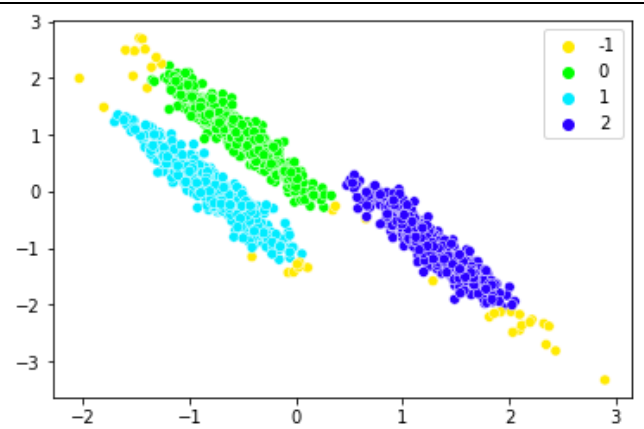


K-means:

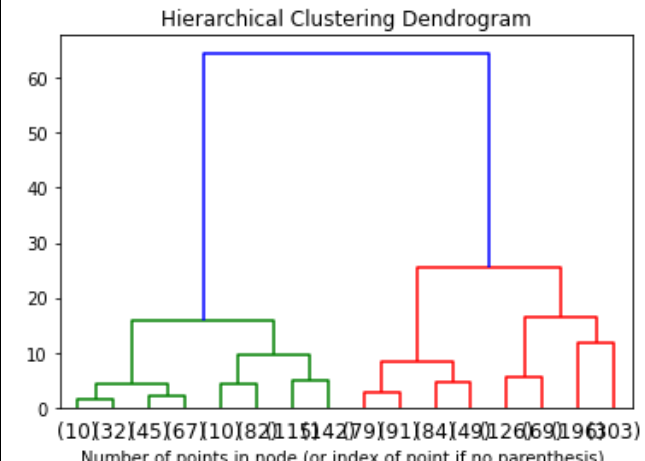
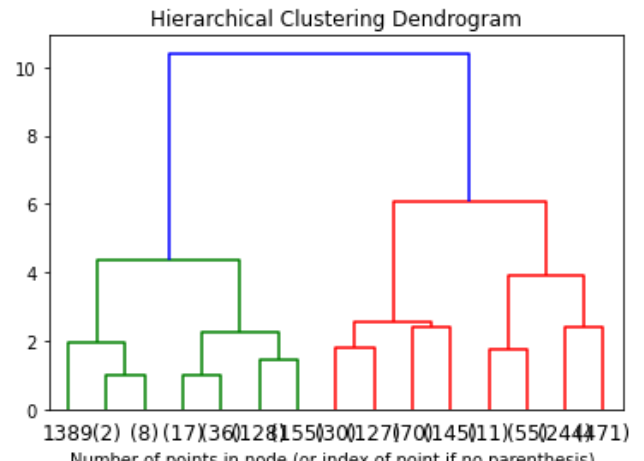


Calinski-Harabasz index: 3637.91297	Calinski-Harabasz index: 2453.17409
Davies-Bouldin index: 0.70025	Davies-Bouldin index: 0.78958
Silhouette Coefficient: 0.50997	Silhouette Coefficient: 0.48512
Cluster_cohesion: 511.92170	Cluster_cohesion: 701.35184
Adjusted Rand Index: 0.6074569	Adjusted Rand Index: 0.3930642

DBSCAN: (желтые точки слева метод определяет как выбросы)

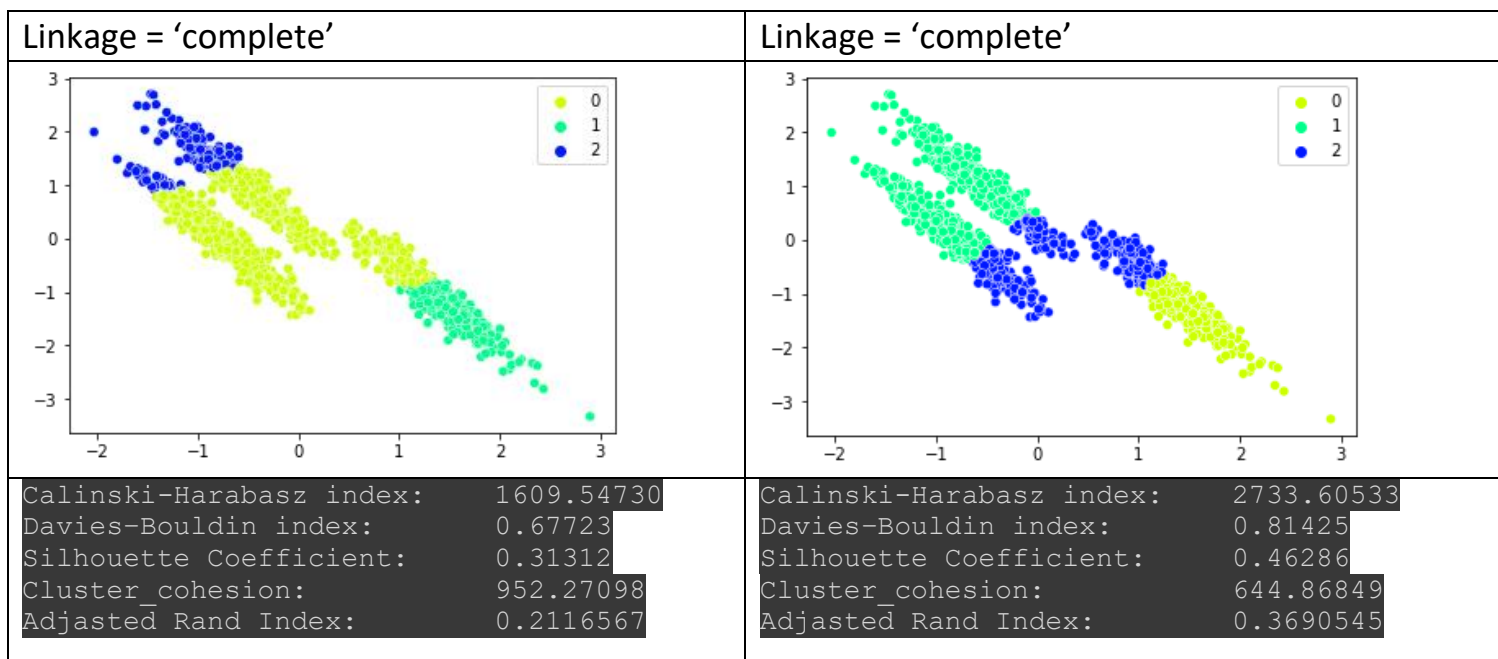
DBSCAN (евклидово расстояние) eps=0.16, min_samples=6	DBSCAN (манхэттенское расстояние) eps=0.2, min_samples=7
	
Calinski-Harabasz index: 1401.44504 Davies-Bouldin index: 2.61179 Silhouette Coefficient: 0.46237 Cluster_cohesion: 569.25084 Adjusted Rand Index: 0.9718918	Calinski-Harabasz index: 1376.77190 Davies-Bouldin index: 2.40096 Silhouette Coefficient: 0.40455 Cluster_cohesion: 547.56401 Adjusted Rand Index: 0.9615839

Дендрограмма:

(евклидово расстояние)	(манхэттенское расстояние)
<p>Hierarchical Clustering Dendrogram</p>  <p>Number of points in node (or index of point if no parenthesis).</p>	<p>Hierarchical Clustering Dendrogram</p>  <p>Number of points in node (or index of point if no parenthesis).</p>

Аггломеративная кластеризация

AgglomerativeClustering (евклидово расстояние)	AgglomerativeClustering (манхэттенское расстояние)
--	--



Датасет 7 - <http://archive.ics.uci.edu/ml/datasets/Abalone>

### Abalon Data Set

Данные представляют собой таблицу измерений 8 физических параметров морских ракушек. По этим данным нужно предсказать возраст ракушки, который считается по числу колец на срезе. Возраст (число колец) здесь является классом. Итого таблица содержит 4417 строк.

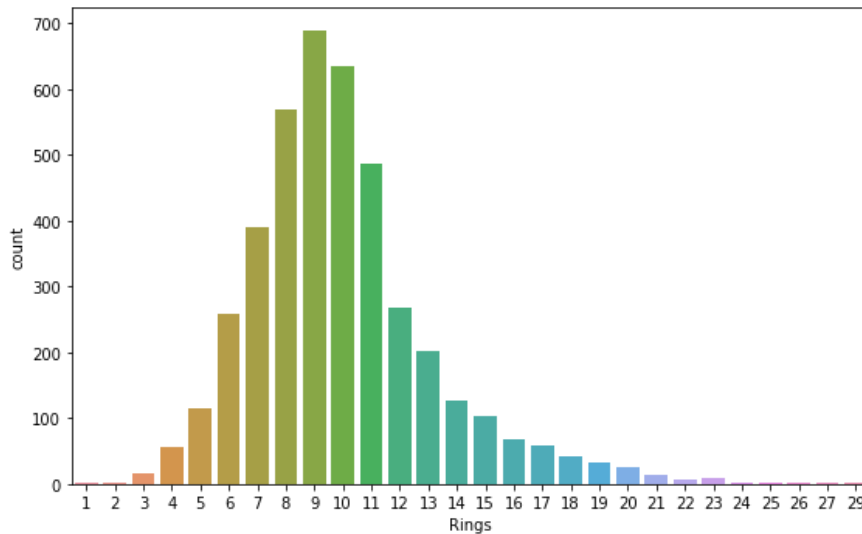
Признаки:

- Sex (категориальный) - M, F, and I (infant)
- Length / Longest shell measurement (числовой)
- Diameter / perpendicular to length (числовой)
- Height / with meat in shell (числовой)
- Whole weight (числовой)
- Shucked weight (числовой)
- Viscera weight (числовой)
- Shell weight (числовой)

Проведем предобработку данных:

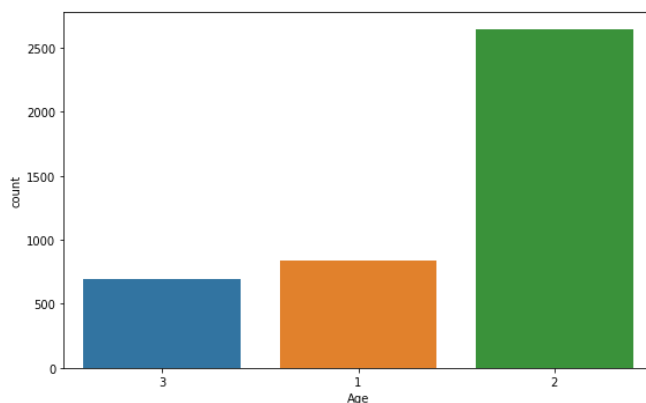
	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

В датасете Abalone нет пропущенных значений, но классе крайне не сбалансированы (пока не будем с этим ничего делать). Ракушек возрастом до 5 и от 16 и более крайне мало.



Чтобы более-менее сделать группы равными по размеру сгруппируем новые классы:

Rings от 0 до 7 – 0 класс, от 8 – 11 – 1 класс, остальные – 2 класс.

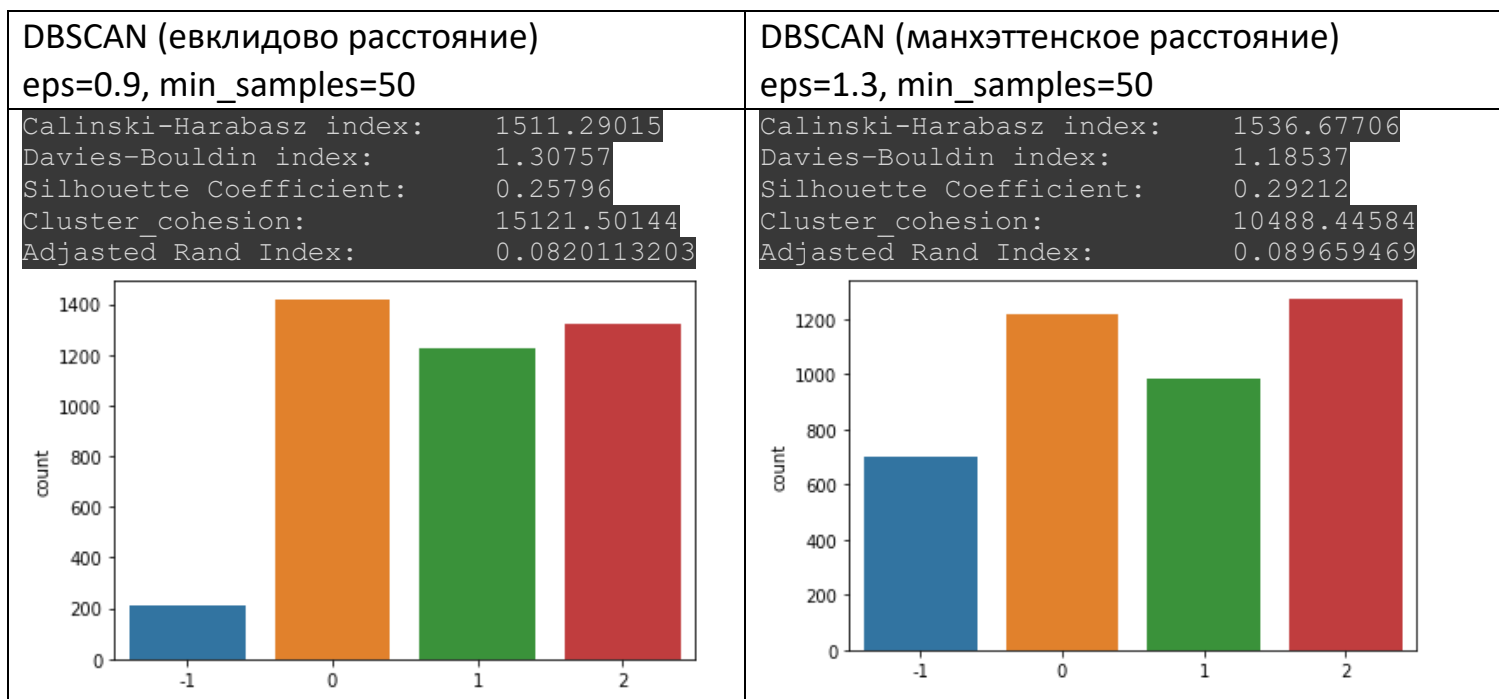


Признак ‘Sex’ категориальный, переведем его в числовое пространство, просто занумеровав (LabelEncoder). Далее данные стандартизуем и можно приступить к кластеризации

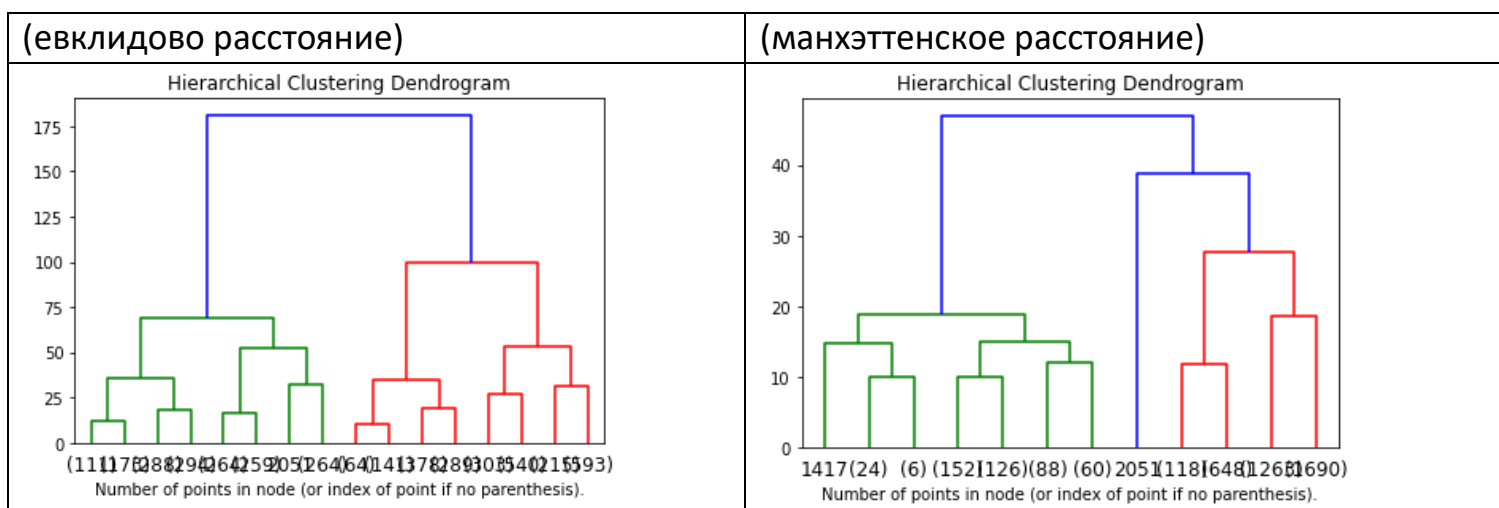
K-means:

K-means (евклидово расстояние)	K-medoids (манхэттенское расстояние)
Calinski-Harabasz index: 4141.03163	Calinski-Harabasz index: 4034.79284
Davies-Bouldin index: 1.01720	Davies-Bouldin index: 1.04693
Silhouette Coefficient: 0.34819	Silhouette Coefficient: 0.41138
Cluster_cohesion: 11197.62971	Cluster_cohesion: 11391.95557
Adjusted Rand Index: 0.146148473	Adjusted Rand Index: 0.138012995

DBSCAN (синие столбики – доля выбросов)



## Дендрограмма:



## Агломеративная кластеризация

AgglomerativeClustering (евклидово расстояние) Linkage = 'ward'	AgglomerativeClustering (манхэттенское расстояние) Linkage = 'complete'
Calinski-Harabasz index: 3707.17758 Davies-Bouldin index: 0.98161 Silhouette Coefficient: 0.32394 Cluster_cohesion: 12036.08122 Adjusted Rand Index: 0.167812161	Calinski-Harabasz index: 867.36840 Davies-Bouldin index: 0.59840 Silhouette Coefficient: 0.38060 Cluster_cohesion: 23605.44877 Adjusted Rand Index: 0.006012446

Датасет 8 - <https://www.kaggle.com/creepyghost/uci-ionosphere>

### UCI Ionosphere Data Set

Эти радиолокационные данные были собраны системой в Goose Bay, Labrador. Эта система состоит из фазовой решетки из 16 высокочастотных антенн с общей передаваемой мощностью порядка 6,4 киловатт. Мишенями служили свободные электроны в ионосфере. "Хорошие" радиолокационные сигналы — это те, которые показывают присутствие некоторого типа структуры в ионосфере. «Плохими» считаются те, которые не возвращаются; их сигналы проходят через ионосферу.

Полученные сигналы обрабатывались с помощью автокорреляционной функции, аргументы которой - время импульса и номер импульса. Для системы Goose Bay было 17 номеров импульсов. Экземпляры в этой базе данных описываются двумя атрибутами на один номер импульса, соответствующими комплексным значениям, возвращаемым функцией в результате сложного электромагнитного сигнала.

Информация об атрибутах:

- Все 34 непрерывные
- 35-й атрибут является либо «хорошим», либо «плохим» в соответствии с приведенным выше определением. Это задача двоичной классификации. Итого таблица содержит 315 строк.

Проведем предобработку данных:

Col_21	Col_22	Col_23	Col_24	Col_25	Col_26	Col_27	Col_28	Col_29	Col_30	Col_31	Col_32	Col_33	Class
0.29674	0.36946	-0.47357	0.56811	-0.51171	0.41078	-0.46168	0.21266	-0.34090	0.42267	-0.54487	0.18641	-0.45300	g
0.45300	-0.18056	-0.35734	-0.20332	-0.26569	-0.20468	-0.18401	-0.19040	-0.11593	-0.16626	-0.06288	-0.13738	-0.02447	b
0.27502	0.43385	-0.12062	0.57528	-0.40220	0.58984	-0.22145	0.43100	-0.17365	0.60436	-0.24180	0.56045	-0.38238	g
0.00000	0.00000	0.00000	1.00000	0.90695	0.51613	1.00000	1.00000	-0.20099	0.25682	1.00000	-0.32382	1.00000	b
0.35575	0.02309	-0.52879	0.03286	-0.65158	0.13290	-0.53206	0.02431	-0.62197	-0.05707	-0.59573	-0.04608	-0.65697	g

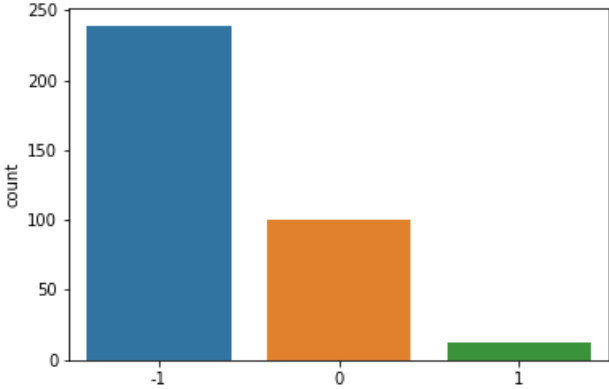
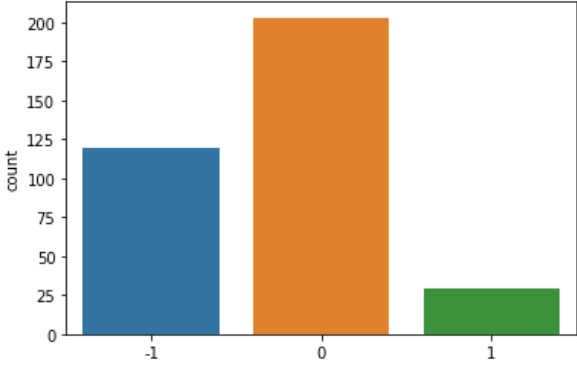
Пропущенных данных нет. Метки класса занумеруем с помощью LabelEncoder, приведем данные к стандартному виду и можно приступить к кластеризации.

K-means:

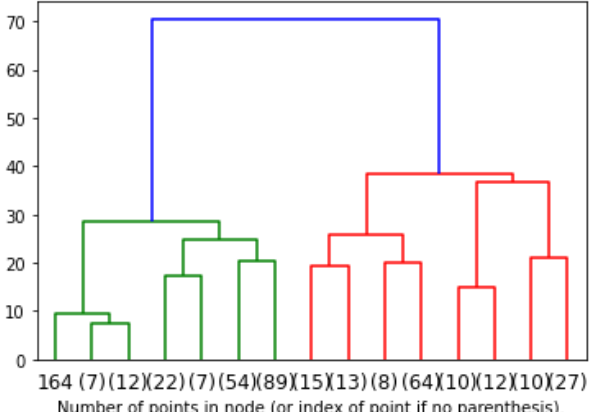
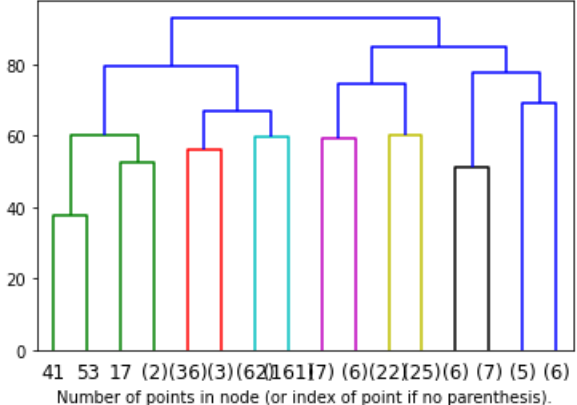
K-means (евклидово расстояние)	K-medoids (манхэттенское расстояние)
Calinski-Harabasz index: 95.91251	Calinski-Harabasz index: 87.84073
Davies-Bouldin index: 1.68190	Davies-Bouldin index: 1.74445
Silhouette Coefficient: 0.27041	Silhouette Coefficient: 0.27808
Cluster cohesion: 9085.98185	Cluster cohesion: 9253.86935
Adjusted Rand Index: 0.16790201	Adjusted Rand Index: 0.08967090



## DBSCAN (синие столбики – доля выбросов)

DBSCAN (евклидово расстояние) eps=1.7, min_samples=12	DBSCAN (манхэттенское расстояние) eps=11, min_samples=5
Calinski-Harabasz index: 118.87515 Davies-Bouldin index: 0.44442 Silhouette Coefficient: 0.59466 Cluster_cohesion: 294.44481 Adjusted Rand Index: 0.7634403	Calinski-Harabasz index: 74.55649 Davies-Bouldin index: 0.97010 Silhouette Coefficient: 0.39940 Cluster_cohesion: 3337.26996 Adjusted Rand Index: -0.0922745
	

## Дендрограмма:

(евклидово расстояние)	(манхэттенское расстояние)
	

## Агломеративная кластеризация

AgglomerativeClustering (евклидово расстояние) Linkage = 'ward'	AgglomerativeClustering (манхэттенское расстояние) Linkage = 'complete'
Calinski-Harabasz index: 95.43817 Davies-Bouldin index: 1.68755 Silhouette Coefficient: 0.27130 Cluster_cohesion: 9095.67921 Adjusted Rand Index: 0.17747908	Calinski-Harabasz index: 17.32822 Davies-Bouldin index: 3.94652 Silhouette Coefficient: 0.23749 Cluster_cohesion: 11035.09577 Adjusted Rand Index: 0.085825800



## Вывод

В данной работе проводилось сравнение трех методов кластеризации с двумя функциями расстояния на разных данных.

В случае когда сгущения точек находятся далеко друг от друга или близко, но не пересекаясь (датасеты 1, 2) то все методы провели кластеризацию правильно.

Если группы накладываются друг на друга (датасеты 3, 4), то каждый метод дает разный результат. Причем k-means и аггломеративный подход дали приблизительно схожее разделение, хотя по adjusted rang index видно, что деление неидеально. DBSCAN совсем плохо смог выделить группы (для манхэттенской метрики вообще либо получались три маленькие группы, а большая часть точек относилась к выбросам, либо вообще не удавалось выделить требуемые 3 группы), с точки зрения метода данные больше похожи на единое облако окруженное большим количеством выбросов.

Зато на датасетах 5, 6 DBSCAN показал лучший результат. Аггломеративный метод справился только на одном из них, k-means -справился плохо. Это связано с данными, так как k-means хорошо работает с данными, которые обладают сферической формой, при этом группы в идеале должны не сильно отличаться в размерах. А DBSCAN хоть и плохо разделяет группы, если они имеют большую площадь пересечения, но он может выделить группы произвольной формы, если правильно задать параметры, и не будет сильных пересечений с соседними кластерами.

В датасете 7 лучше всего сработали k-means и аггломеративный метод, но по метрикам качества видно, что данные плохо кластеризировались. В датасете 8 по adjusted rang index хорошо показал себя DBSCAN, но также он очень почти треть данных относит к выбросам.

В итоге, видно, что каждый метод может хорошо работать на тех или иных данных, так как в основе каждого лежат разные гипотезы о том, какие точки считать членами одного кластера. Также влияние оказывает и метрика (хотя в данной работе почти всегда результат получался выше при использовании расстояния Евклида). Если бы у нас не было меток классов изначально, и мы могли бы использовать только внутренние показатели качества, то оценивать проводимую кластеризацию было бы еще сложнее, так как и в этих метриках тоже в основе лежат разные предположения о структуре данных, что требует дополнительного анализа от исследователя.

## Инструкция по запуску

Для запуска потребуется использовать Anaconda (Jupyter Notebook), но предпочтительнее запускать из Google Colab (<https://colab.research.google.com/drive/1F8MN28nnWvQGzfPqfJGUOWtFU04vE9qM?usp=sharing>), так как там уже гарантированно установлены используемые библиотеки. Также все датасеты, использованные в отчете, должны лежать в одной папке с кодом, иначе нужно будет изменить в коде путь к ним.

Затем нужно просто последовательно выполнить код в ячейках. Никакие параметры менять не нужно, иначе не получится воспроизвести достигнутые результаты классификации в точности как полученные в отчете к работе.