**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра МОЭВМ**

отчет

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

Тема: **Предобработка данных.**

|  |  |  |
| --- | --- | --- |
| Студент гр. 6302 |  | Барбарич И.Г. |
| Руководитель |  | Жангиров Т. Р. |

Санкт-Петербург

2020

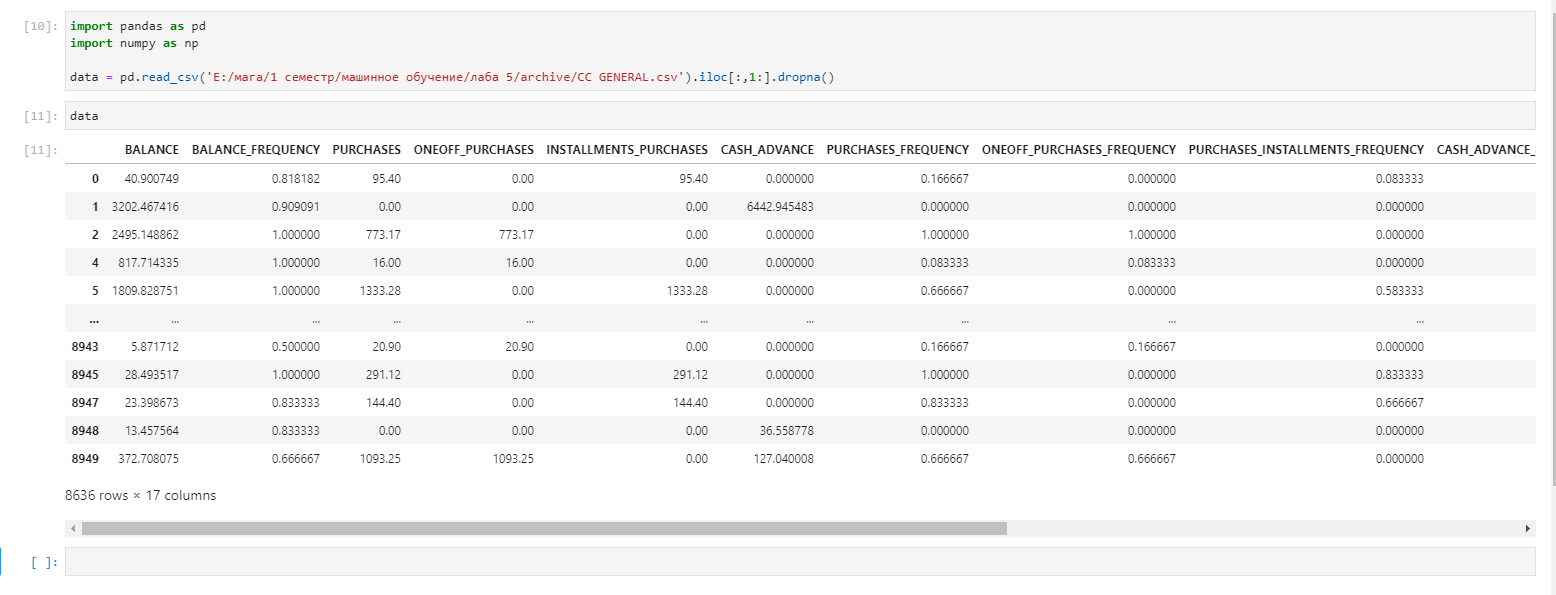
**Цель работы**

Ознакомиться с методами кластеризации модуля Sklearn

**Загрузка данных:**

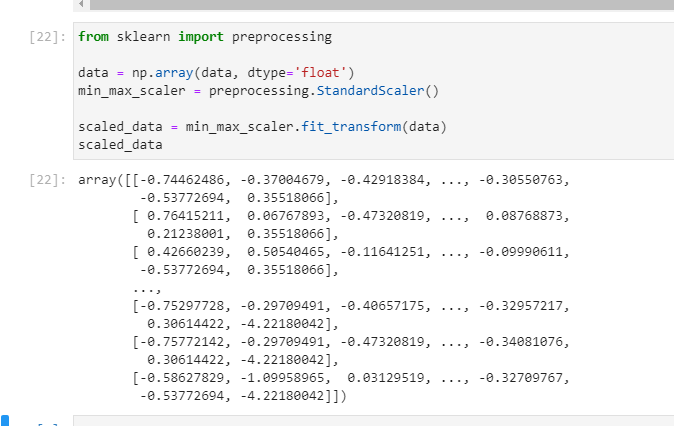
1. Загрузить датасет по ссылке: https://www.kaggle.com/arjunbhasin2013/ccdata . Данные представлены в виде csv файла. Датасет содержит пропущенные значения

2. Создать Python скрипт. Загрузить данные в датафрейм, убрав столбец с метками и откинув наблюдения с пропущенными значениями



**DBSCAN**

Так как разные признаки лежат в разных шкалах, то стандартизируем данные



Проведем кластеризацию методов DBSCAN при параметрах по умолчанию. Выведем метки кластеров, количество кластеров, а также процент наблюдений, которые кластеризовать не удалось



Опишите все параметры, которые принимает DBSCAN:  
Eps (float), default=0.5 - Максимальное расстояние между двумя образцами для того, чтобы один рассматривался как находящийся в окрестности другого.

min\_samples (int), default=5 - Число выборок (или общий вес) в окрестности точки, рассматриваемой в качестве основной точки. Это включает в себя и саму точку.

Metric (string), or callable, default=’euclidean’ - Метрика, используемая при вычислении расстояния между экземплярами в массиве объектов. Если метрика является строкой или вызываемой, она должна быть одной из опций, разрешенных

metric\_params (dict), default=None - Дополнительные аргументы ключевого слова для функции метрики.

Additional keyword arguments for the metric function.

algorithm{‘auto’, ‘ball\_tree’, ‘kd\_tree’, ‘brute’}, default=’auto’ - Алгоритм, который будет использоваться для вычисления точечных расстояний и поиска ближайших соседей.

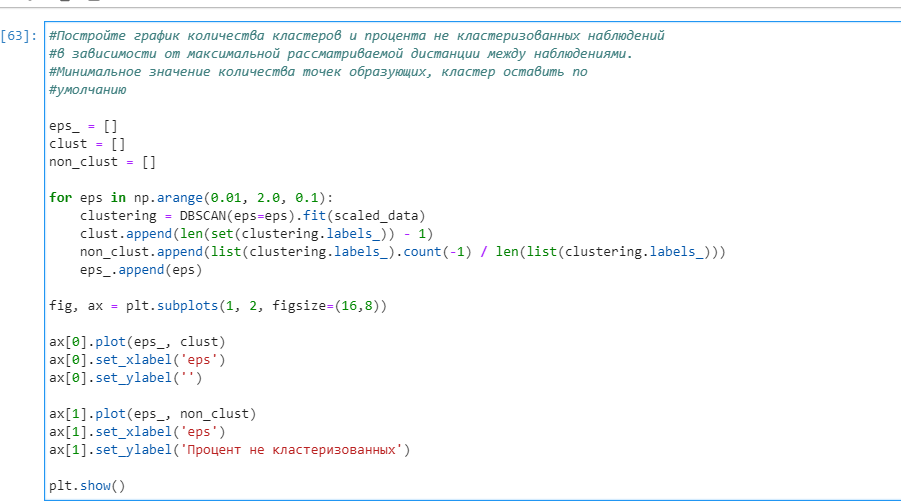
leaf\_size (int), default =30 - Размер листа передается в BallTree или cKDTree. Это может повлиять на скорость построения и запроса, а также на объем памяти, необходимый для хранения дерева. Оптимальное значение зависит от характера задачи.

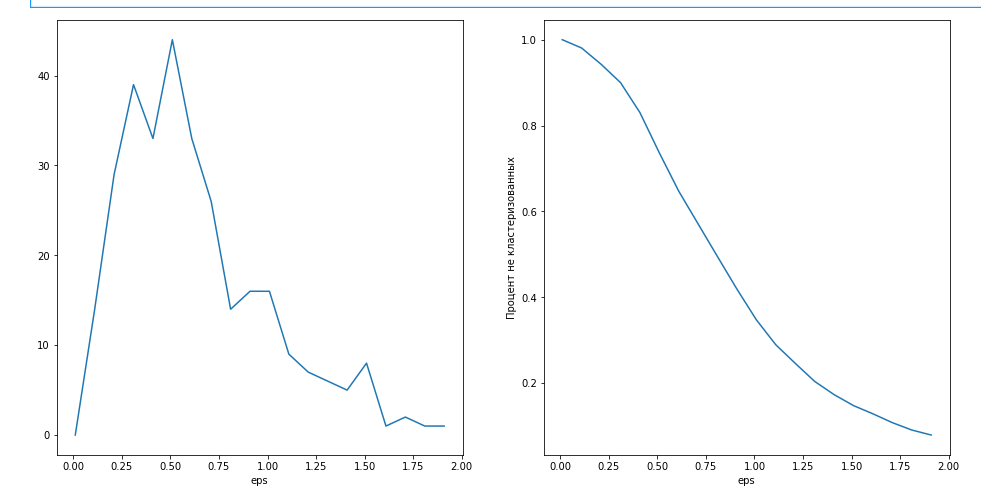
pfloat, default=None - Мощность метрики Минковского, которая будет использоваться для вычисления расстояния между точками. Если нет, то p=2 (эквивалентно евклидову расстоянию).

n\_jobsint, default=None

Количество выполняемых параллельных заданий. None означает 1, если только в контексте joblib.parallel\_backend. -1 означает использование всех процессоров.

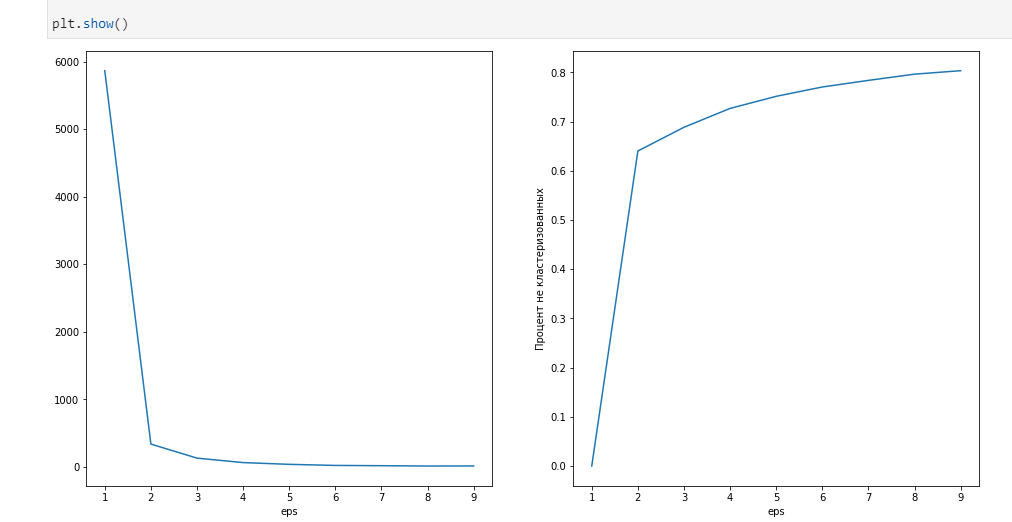
4. Постройте график количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной рассматриваемой дистанции между наблюдениями. Минимальное значение количества точек образующих, кластер оставить по умолчанию





5. Постройте график количества кластеров и процента не кластеризованных наблюдений в зависимости от минимального значения количества точек, образующих кластер. Максимальную рассматриваемую дистанцию между наблюдениями оставьте по умолчанию

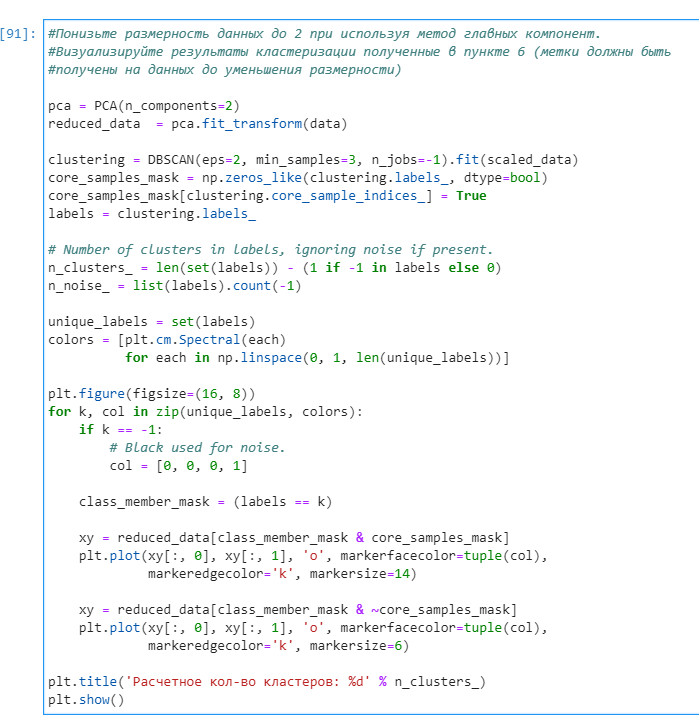


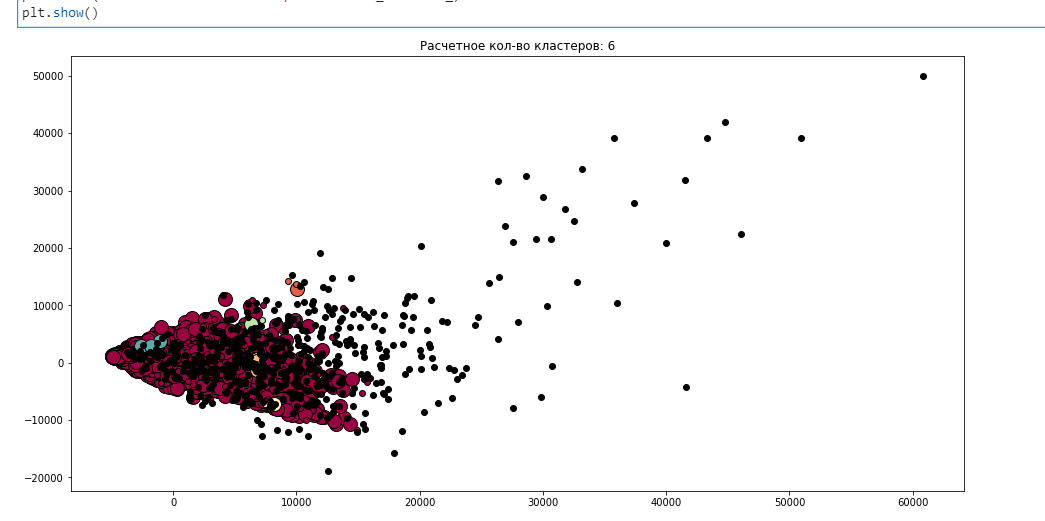


6. Определите значения параметров, при котором количество кластеров получается от 5 до 7, и процент не кластеризованных наблюдений не превышает 12%.



7. Понизьте размерность данных до 2 при используя метод главных компонент. Визуализируйте результаты кластеризации полученные в пункте 6 (метки должны быть получены на данных до уменьшения размерности). гайд по визуализации





**OPTICS**

Опишите параметры метода OPTICS, а также какими атрибутами он обладает

Параметры:

min\_samples (int > 1 или float между 0 и 1, по умолчанию=5) - Число выборок в окрестности точки, рассматриваемой в качестве базовой точки.

max\_eps (float, default=np.Inf) - Максимальное расстояние между двумя образцами для того, чтобы один из них рассматривался как находящийся в окрестности другого.

metric str or callable, по умолчанию= ' minkowski’ - Метрика, используемая для вычисления расстояния.

P (int), по умолчанию=2 - gараметр для метрики

metric\_paramsdict, default=None - Дополнительные аргументы ключевых слов для метрической функции.

cluster\_method (str), default= ' xi’ - Метод экстракции используется для извлечения кластеров с использованием расчетной достижимости и упорядоченности. Возможные значения “ "xi”и "dbscan".

EPS (float), по умолчанию=нет - Максимальное расстояние между двумя образцами для того, чтобы один из них рассматривался как находящийся в окрестности другого.

Xi (float) между 0 и 1, по умолчанию=0,05 - Определяет минимальную крутизну на участке достижимости , образующем границу кластера.

min\_cluster\_size (int > 1 или float между 0 и 1), default=None - Минимальное число образцов в оптическом кластере, выраженное как абсолютное число или доля от числа образцов (округленное до значения не менее 2).

leaf\_sizeint, по умолчанию=30 - Размер листа передается в BallTreeОр KDTree. Это может повлиять на скорость построения и запроса, а также на объем памяти , необходимый для хранения дерева.

n\_jobsint, default=None - Количество параллельных заданий для выполнения поиска соседей. None означает 1, если только в [joblib.parallel\_backend](https://joblib.readthedocs.io/en/latest/parallel.html#joblib.parallel_backend)контексте. -1 означает использование всех процессоров.

Атрибуты

labels\_ - ndarray of shape (n\_samples,) - Метки кластеров для каждой точки набора данных задаются функцией fit ().

ordering\_ - ndarray of shape (n\_samples,) - Кластер упорядочил список выборочных индексов.

core\_distances\_ - ndarray of shape (n\_samples,) - Расстояние, на котором каждый образец становится базовой точкой, индексируется по порядку объектов.

cluster\_hierarchy\_ - ndarray of shape (n\_clusters, 2) - Список кластеров в виде [start, end] в каждой строке, со всеми индексами включительно.

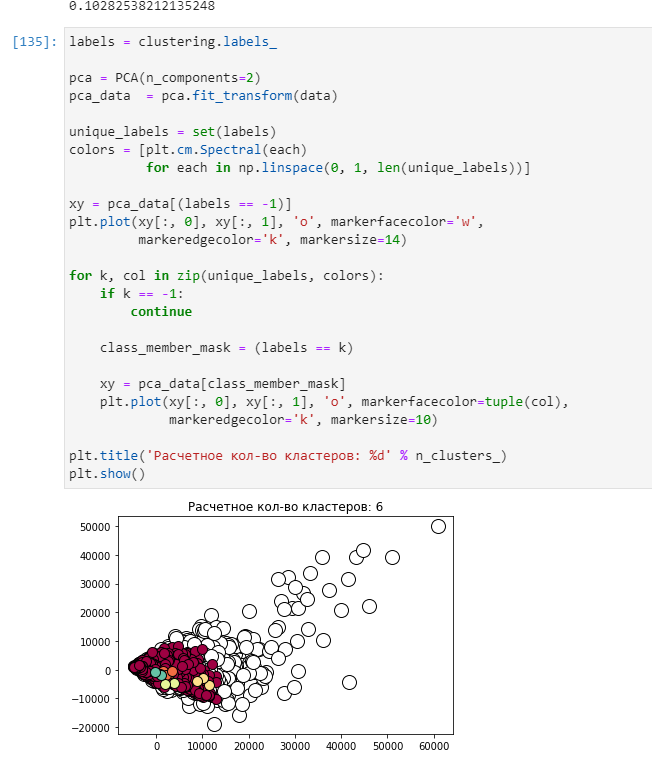
Найдите такие параметры метода OPTICS (\*max\_eps \*и min\_samples) при которых, чтобы получить результаты близкие к результатам DBSCAN из пункта 6

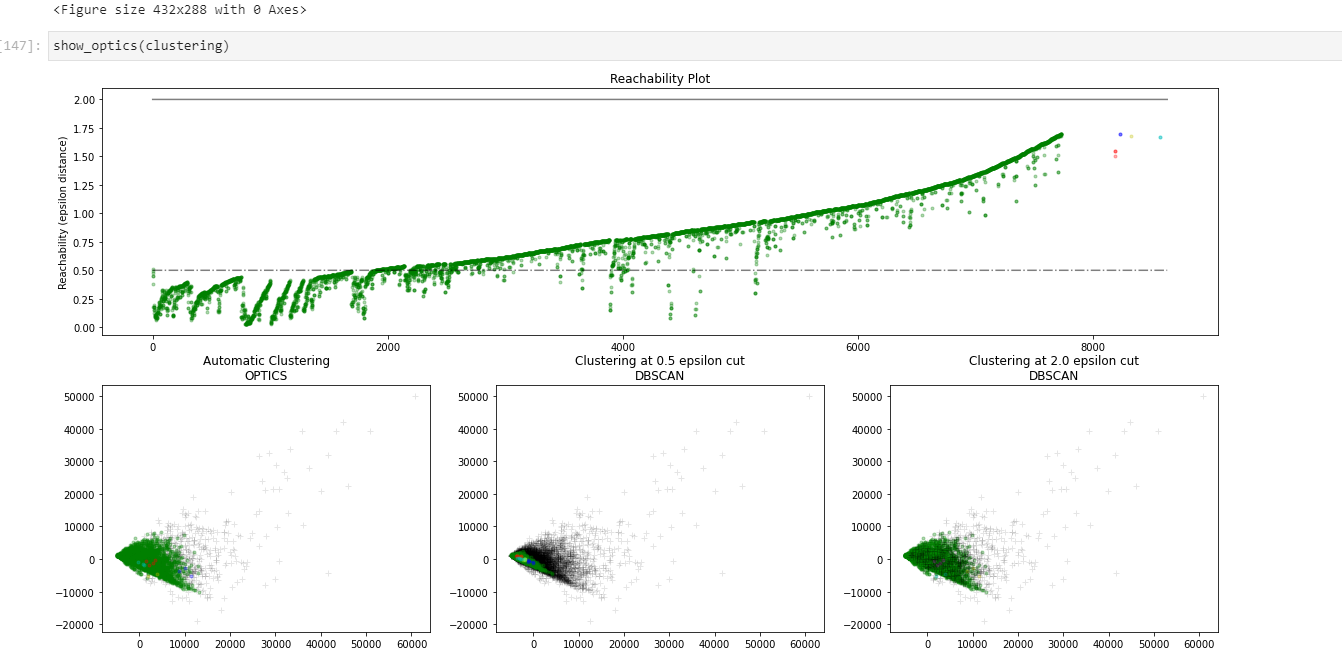


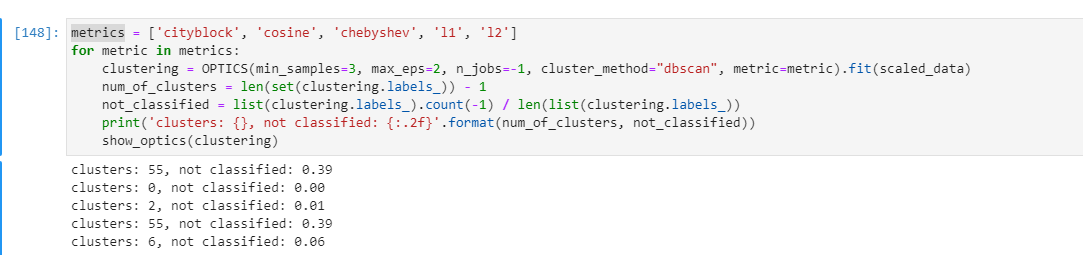
В чем отличия от метода OPTICS от метода DBSCAN

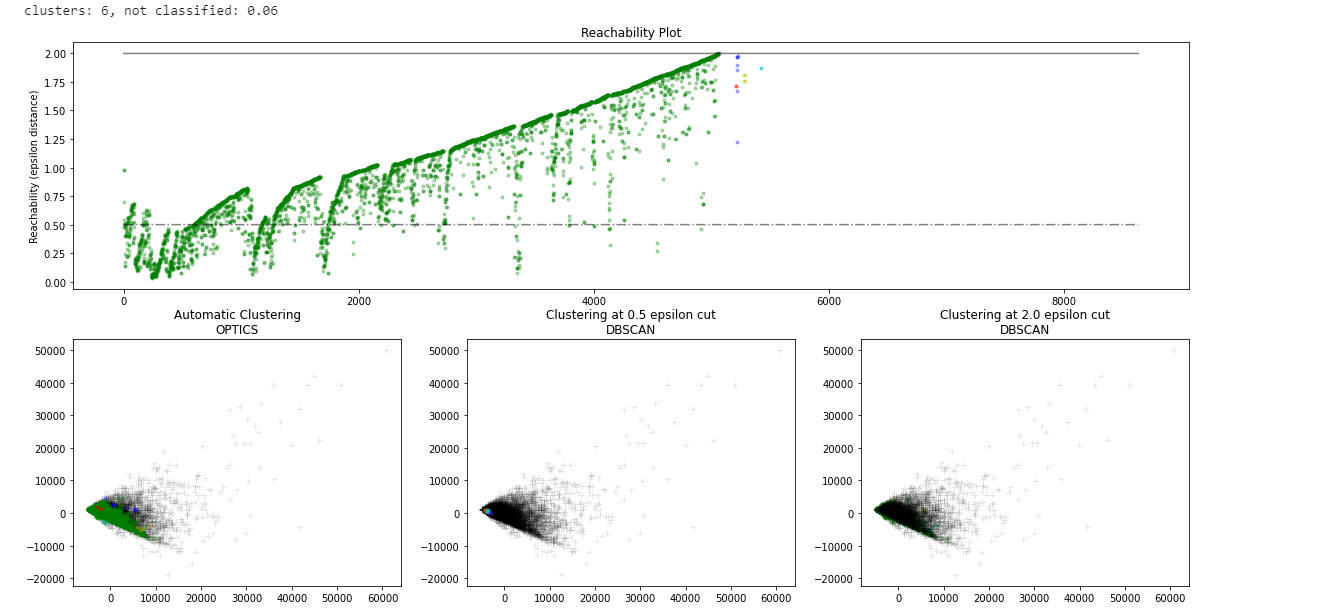
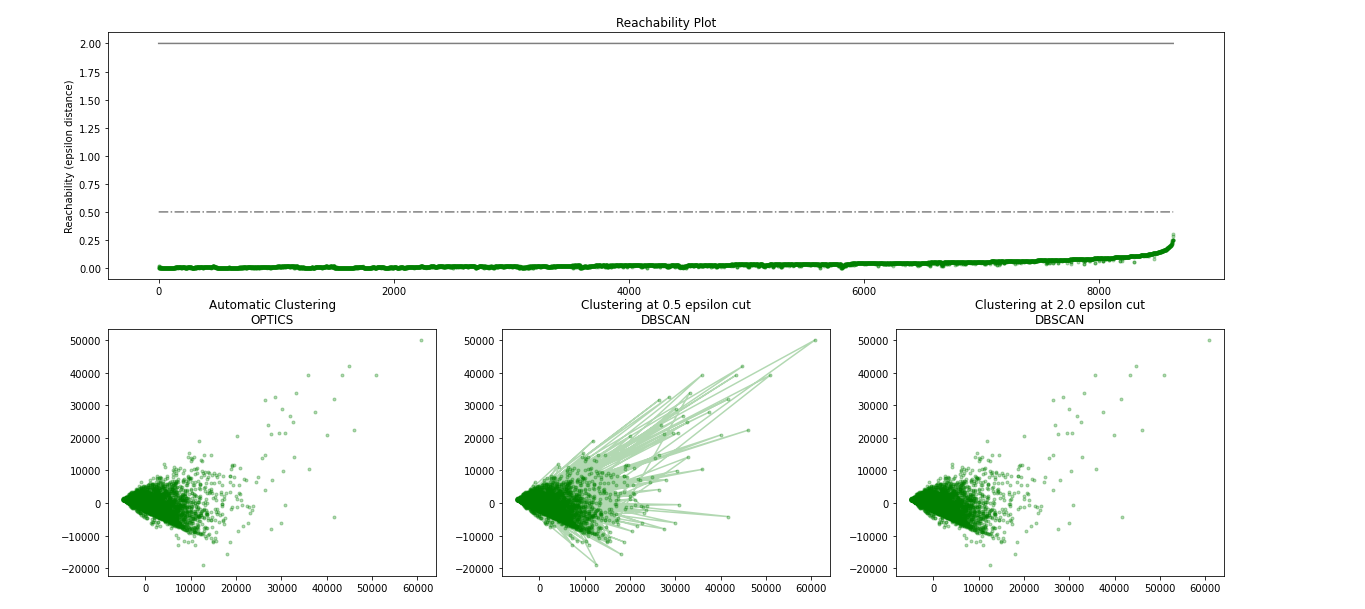
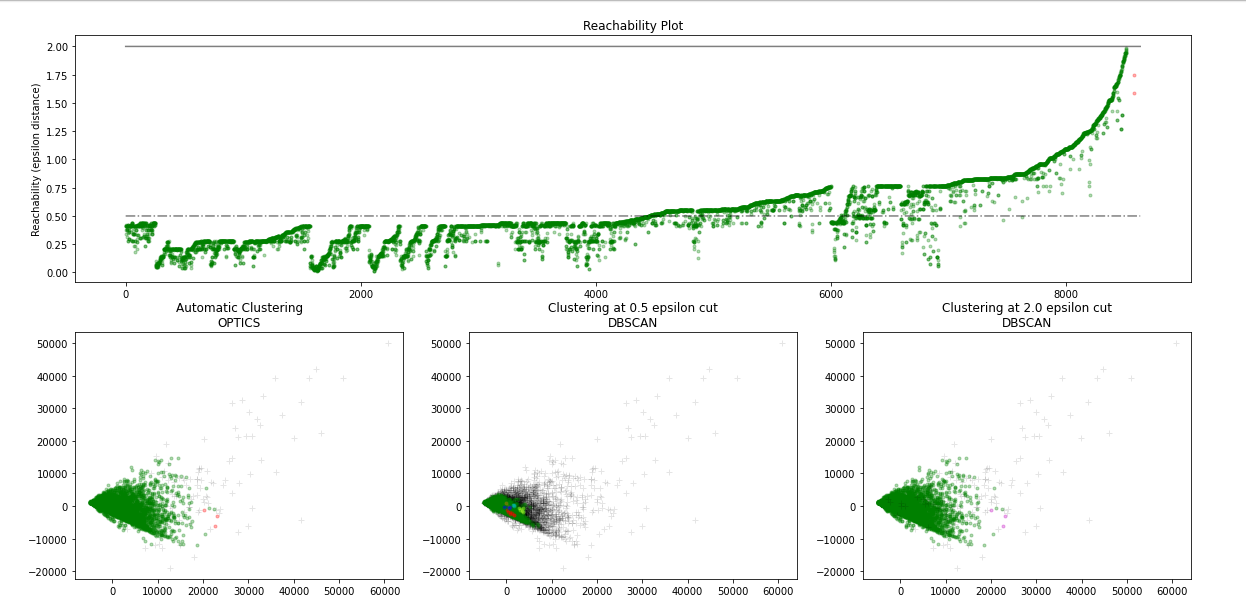
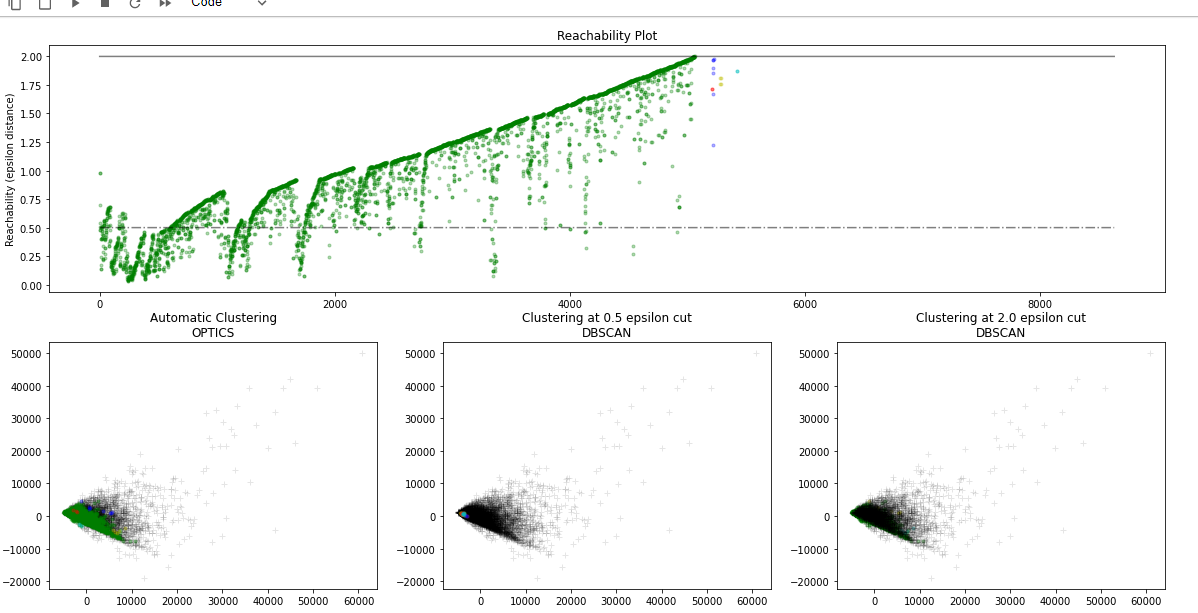
В Optics параметр eps не учитывается, если и учитывается, то только задает максимальное значение, в отличие от DBSCAN, где значение eps является важным параметром.

3. Визуализируйте полученный результат, а также постройге график достижимости (reachable plot) гайд



4. Исследуйте работу  метода OPTICS с использованием различных метрик (выберите не менее 5 метрик)



1. 
2. 
3. 
4. 
5. 