|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра ИУ5 «Системы обработки информации и управления»**

Отчет по рубежному контролю №2

Кластеризация данных.

по дисциплине «Машинное обучение»

Выполнил:

студент группы ИУ5-63

Кондратьев Максим

31.05.2019

Проверил:

к.т.н., доц., Ю. Е. Гапанюк

подпись, дата

Москва, 2019 г.

Задача:

Необходимо решить задачу кластеризации на основе любого выбранного Вами датасета.

Кластеризуйте данные с помощью трех различных алгоритмов кластеризации. Алгоритмы выбираются произвольным образом, рекомендуется использовать алгоритмы из лекции.

Сравните качество кластеризации для трех алгоритмов с помощью следующих метрик качества кластеризации:

Adjusted Rand index

Adjusted Mutual Information

Homogeneity, completeness, V-measure

Коэффициент силуэта

Сделайте выводы о том, какой алгоритм осуществляет более качественную кластеризацию на Вашем наборе данных.

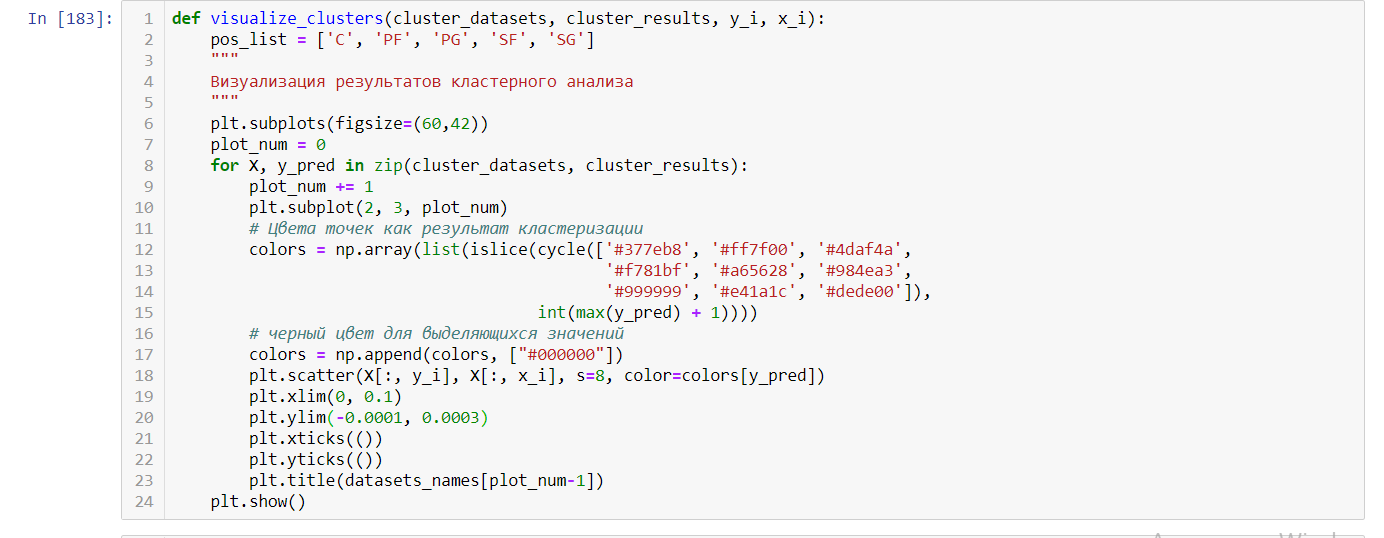
Датасет:

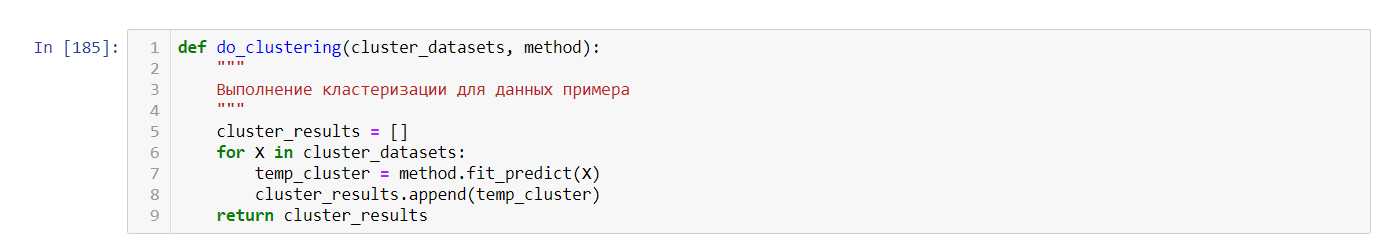
https://www.kaggle.com/drgilermo/nba-players-stats - статистика игроков за 67 сезонов NBA.

Для разработки мне понадобятся ряд модулей и несколько реализованных мной функций для предварительной обработки датасета:



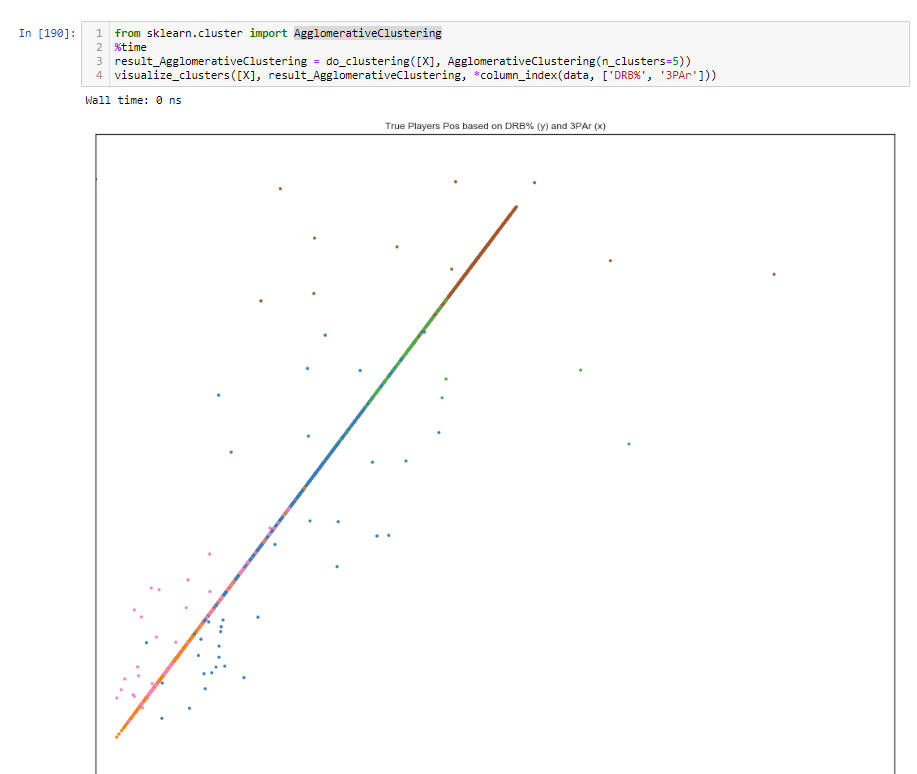
Также я воспользуюсь функциями, описанными в материалах лекции по кластеризации:





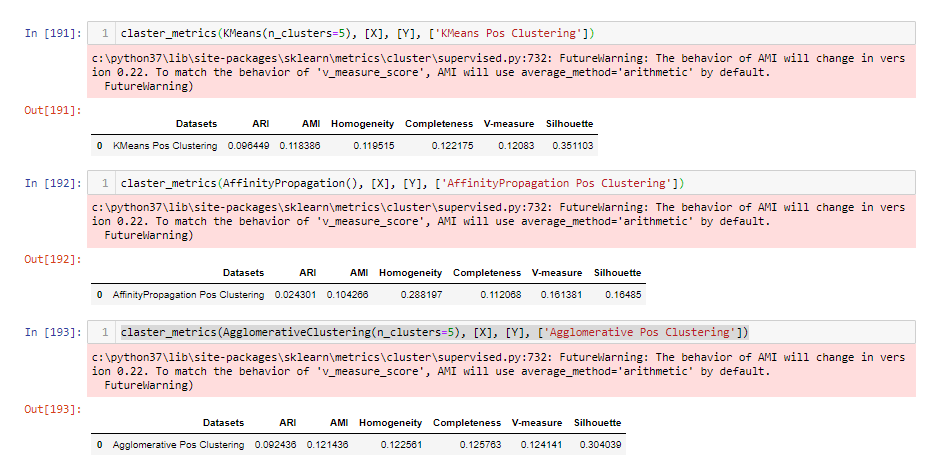
В качестве алгоритмов кластеризации я выбрал KMeans, AffinityPropagation, AgglomerativeClustering. Точное значения количества кластеров известно – их пять, согласно количеству основных позиций игроков в баскетбол. С использованием данных моделей я получил следующие результаты:

|  |  |
| --- | --- |
|  |  |



Метрики для всех моделей:





С учетом результатов, я считаю, что данные слабо подвержены кластеризации. На основе достоверных сведений о принадлежности данных к кластерам, я считаю, что лучшей метрикой в данном случае следует считать Homogeneityили V-measureввиду сильной зашумленности данных, распределенных на 5 кластеров. Отсюда я делаю вывод, что лучше всех справился со своей задачей метод AffinityPropagation.