2 Лабораторная работа

Цель - познакомиться с методом кластеризации K-mean с помощью пакета sklearn.

Для выполнения задания используется набор данных из предыдущей лр.

Что нужно сделать:

1. К текущему датасету необходимо добавить новый аттрибут, например, если есть дата рождения, то посчитать возраст или определить популярность фильма по количеству просмотров и т.п.

2. "Причесать" датасет: удалить выбросы и дубли, обработать пропущенные значения, найти кривые данные и т.п.

3. Построить графики зависимости одной переменной от другой (plt.scatter), описать какие кластеры (группы) вы видите. Построить несколько разных графиков, найти ярковыраженные группы.

4. Изучить перечисленные ноутбуки, по примеру, попробовать применить метод KMeans к вашему датасету (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)

https://www.kaggle.com/kushal1996/customer-segmentation-k-means-analysis

https://www.kaggle.com/karthickaravindan/k-means-clustering-project

https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering

https://www.kaggle.com/sirpunch/k-means-clustering

Если ваш датасет не подходит для задачи, его можно поменять.

5. Написать выводы

Если не установлен sklearn - https://scikit-learn.org/stable/install.html