Задание 5. Предобработка данных и PCA

Необходимо провести анализ датасета (из задания 6) и сделать обработку данных по предложенному алгоритму. Код подготовить в виде файлов *.py и сделать отчет в виде ноутбука с описанием процесса анализа.

Ответить на следующие вопросы:

1. Сколько в датасете объектов и признаков? Дать описание каждому признаку, если оно есть.
2. Сколько категориальных признаков, какие?
3. Столбец с макимальным количеством уникальных значений категориального признака?
4. Есть ли бинарные признаки?
5. Какие числовые признаки?
6. Есть ли пропуски?
7. Сколько объектов с пропусками?
8. Столбец с максимальным количеством пропусков?
9. Есть ли на ваш взгляд выбросы, аномальные значения?
10. Столбец с максимальным средним значением после нормировки признаков через стандартное отклонение?
11. Столбец с целевым признаком?
12. Сколько объектов попадает в тренировочную выборку при использовании train_test_split с параметрами test_size = 0.3, random_state = 42?
13. Между какими признаками наблюдается линейная зависимость (корреляция)?
14. Сколько признаков достаточно для объяснения 90% дисперсии после применения метода PCA?
15. Какой признак вносит наибольший вклад в первую компоненту?

* Доп. задание: построить двухмерное представление данных с помощью алгоритма t-SNE. На сколько кластеров визуально на ваш взгляд разделяется выборка?