**НИУ ВШЭ НН. Факультет ИМиКН. Методы анализа данных.**

**Charge de cours: В.А. Калягин.**

**Домашнее задание 1:** данные,пропуски, выбросы,методы снижения размерности (PCA, MDS, UMAP, …)

Провести необходимые вычисления. Полученные результаты снабдить подробным комментарием.

**Задача.** В Excel таблице собраны персональные данные клиентов банка. Столбцы таблицы отражают следующие сведения о клиенте

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 |
| Возраст (лет) | Количество иждивенцев | Доход (руб./мес.) | Срок проживания в регионе (лет) | Оценка автомобиля заемщика (USD) | Размер кредита  (тыс. руб.) |

**Данные**. **Выполнить первичную обработку данных. Сделать первые выводы.**

1. Рассмотрите данные и вычислите основные характеристики признаков: среднее, стандартное отклонение, медиана, квартили. Постройте гистограммы признаков и блок диаграммы признаков. Сделайте первые выводы о возможном распределении признаков.

2. В данных есть пропуски. Подберите из литературы метод восстановления пропущенных данных и заполните пропуски по этому методу.

3. В данных возможны выбросы. Подберите из литературы метод определения выбросов и удалите объекты, идентифицированные как выбросы по этому методу.

**Снижение размерности. Метод PCA** (Principal Component Analysis=метод главных компонент).

4. Выполнить центрирование данных и такую нормировку, что евклидова норма каждого признака равна 1 (полученная матрица обозначается через Z).

5. Найти SVD разложение матрицы Z (матрицы U, S, V).

6. Исследуйте, сколько скрытых факторов (главные факторы) достаточно для хорошей аппроксимации матрицы данных Z.

7. Замените значения исходных признаков линейными комбинациями главных факторов (матрица Zk, k – число главных факторов). Вычислите норму разности (погрешность аппроксимации) в операторной матричной норме и норме Фробениуса.

8. Используя выделенные главные факторы, найти представления клиентов векторами меньшей размерности (размерности k) и построить карту клиентов. Сделать выводы о клиентах банка.

**Снижение размерности. Метод MDS** (Multidimensional Scaling = многомерное шкалирование)

9. Вычислите расстояние между клиентами любым понравившемся вам способом (расстояние Чебышева, Манхэттенских кварталов, Минковского). Проверьте, можно ли клиентов банка поместить в евклидово пространство размерности 2 с сохранением всех расстояний между клиентами.

10. Используйте любой из алгоритмов MDS для размещения объектов в евклидовом пространстве размерности 2. Оцените погрешность такого размещения. Постройте карту клиентов. Сравните с картой клиентов из п.8 и сделайте выводы.

**Бонус.** **Методы снижени размерности.**

Подберите в литературе метод снижения размерности (UMAP, t-SNE, …) и постройте карту клиентов этим методом. Сравните с картой клиентов построенной по методу РСА.