**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)

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

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

**Снижение размерности. Интерпретация скрытых факторов.**

8. Выписать матрицу коэффициентов разложения признаков zj по главным факторам. Исследовать структуру этой матрицы с целью выделения групп признаков тесно связанных с каждым из главных факторов.

9. На основе анализа структуры матрицы коэффициентов разложения, предложить интерпретацию выделенным главным факторам. Дать интерпретацию построенной выше карте клиентов.

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

10. Предложите способ вычисления расстояния между клиентами и вычислите матрицу расстояний.

11. Проверьте, можно ли клиентов банка поместить в евклидово пространство размерности 2 с сохранением расстояний между клиентами.

12. 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.7 и сделайте выводы.

**Вариант - 7**