Homework_1_МИСОБД

Клиентская аналитика

⚡ Дата выдачи: 03.12.2025
• ⏰ Мягкий дедлайн — 23:59 MSK 23.12.2025 • ⏰ Жёсткий дедлайн — 23:59 MSK 30.01.2026

После мягкого дедлайна будет сгорать оценка — по 0,5 балла за каждый день. После жёсткого дедлайна работы не принимаются без уважительной причины.

Система оценки

Результат решения оценивается по 10-балльной шкале, где
• «9-10» — задание решено полностью, все указанные требования по 4 частям выполнены. • Работа выполнена чисто и аккуратно, без значимых замечаний;
• Проведены дополнительные исследования и применены методы сегментации;
• Сделана презентация с сопроводительным письмом.

• «8» — задание решено полностью, выполнены все 4 части домашней работы: • Проведён анализ данных, предоставлен рабочий код и таблицы по исследованию данных;
• Построены сегментации двумя методами с описанием определения количества кластеров;
• Представлены понятные выводы с подтверждёнными данными (таблицы, графики).

• «6-7» — задание решено не полностью или с недочётами: • Проведён анализ данных, предоставлен рабочий код и таблицы по исследованию данных;
• Построена сегментация хотя бы одним методом;
• Представлены понятные выводы с подтверждёнными данными (таблицы, графики).

• «4-5» — задание решено с существенными недочётами: • Проведён анализ данных, предоставлен рабочий код и таблицы по исследованию данных;
• Выявлены верхнеуровневые зависимости и закономерности по клиентам без построения модели сегментации.

• «0-3» — задание не решено или решено неверно.

⚡ **Два одинаковых или почти одинаковых присланных кода с выводами будут оштрафованы:**
0 баллов обоим студентам за задание, независимо от результатов.

Задача

Вам представлены тестовые данные по заявкам на кредит в банк.
Таблица содержит данные по заявке на кредит, а также подтянутые данные по существующим клиентам банка (транзакционная активность и выданные кредиты).

Вариант указан в гугл-доке напротив вашей фамилии во вкладке "HW_1":
https://docs.google.com/spreadsheets/d/1jiLdR3cYMT5TFsGpF7Xm8HsFnwO2kfU3UF8y6PCWkXM/edit?gid=424296667#gid=424296667

Варианты и описание данных представлены в папке:
https://disk.yandex.ru/d/i0z880-cKcTSzQ

Задание состоит из 4 частей:

1. Исследование данных и обработка данных для проведения последующей сегментации. (Вес задания 30%)

2. Сделать минимум 2 метода сегментации (Вес задания 30%)

3. Провести оценку по определению количества кластеров — минимум 5 кластеров (Вес задания 20%)

4. Составить профили клиентов на основе 2 методов сегментации (Вес задания 20%)

⚡ Все комментарии, описания сегментов и выводы по домашней работе можно делать сразу в тетрадке кода.

Описание задания

1. Исследование данных и обработка данных для проведения последующей сегментации. (Вес задания 30%)

Исследуем распределения по данным:

- ✅ Рассчитываем количество уникальных значений, нулевых и пустых значений + доля в % от общего количества;
- ✅ Среднее значение, медиана, стандартное отклонение, минимум, максимум, тип данных по каждому показателю в предоставленных данных;
- ✅ Исследуем распределение данных по полу, возрасту и другим категориальным показателям;

Делаем проверку на:

- ✅ Полноту данных по клиентам;

- ✅ Пропущенные и нулевые значения в полях;
- ✅ Наличие некорректных знаков;

Готовим итоговую витрину данных для сегментации, при необходимости:

- ✅ Корректируем данные – исправляем ошибки; исключаем клиентов с большим числом пропусков или восстанавливаем пропущенные значения;
- ✅ Переводим категориальные показатели в целочисленные; Описываем все пояснения по исследованию данных и по всем преобразованиям данных и графикам.

Результат 1 части:

- ✅ Таблицы, сводные отчёты и графики (например, гистограммы и диаграммы Бокса и Вискера);
- ✅ Все данные должны быть сопровождены выводами по полученному результату исследования (3-4 предложения);
- ✅ Финальная витрина для построения сегментации для 2 части.

Снижение оценки:

- Много таблиц и графиков без пояснений;
- Недочистили данные;
- Не убрали неклиентов банка из финальной витрины;
- Сводные таблицы и графики без учёта сортировки;

2. Сделать минимум 2 метода сегментации (Вес задания 30%)

Примеры сегментаций:

- Бизнес-правила;
- Квантили (RFM);
- Кластеризация с учителем (Дерево решений, регрессия, нейросети, градиентный бустинг и др.);
- Кластеризация без учителя (Метод K-средних, EM алгоритм, градиентный спуск и др.)

⚡ Вне зависимости от выбранного метода сегментации, выделение групп клиентов должно соответствовать следующим условиям:

- Внутри сегмента однородность максимальная;
- Между сегментами однородность минимальная;

Результат 2 части:

- ✅ Сделаны 2 сегментации;
- ✅ Минимальное количество кластеров — 5;

Снижение оценки:

- Много таблиц и графиков без пояснений;

- Меньше 5 кластеров по сегментации;
- Один из кластеров больше 50%;

3. Провести оценку по определению количества кластеров — минимум 5 кластеров (Вес задания 20%)

Если выбрали сегментацию:

- Бизнес-правила — сводники минимум с 3 группировками + выводы, почему выбрали именно эти параметры для группировки;
- Квантили (RFM) — можно сделать по 2-3 параметрам, по итогу сегменты должны быть сгруппированы в группы (отток, VIP, core и т.д.);
- Кластеризация без учителя — минимум 3 метода определения количества кластеров (метод локтя, Калински-Харабаз, Дарбина-Уотсона, Дэвиса-Болдина, метод силуэтов);
- Кластеризация с учителем — минимум 3 метрики качества (accuracy, ROC-AUC, F1, Gini);

Результат 3 части:

- ✅ Сделана оценка по определению количества кластеров несколькими методами;
- ✅ Понятные графики/таблицы, где видно, почему именно такое количество кластеров вы выбрали;
- ✅ Сделать круговую диаграмму с % и количеством;

Снижение оценки:

- Нет выводов по определению количества кластеров;
- Нет графиков/таблиц, подтверждающих выбор количества кластеров;

4. Составить профили клиентов на основе 2 методов сегментации (Вес задания 20%)

Результат 4 части:

- ✅ Сформированы портреты клиентов банка на основе полученных данных по сегментациям;
- ✅ Дана интерпретация по каждому полученному сегменту;

Пример описания профилей (неэталонный, у вас может быть иная визуализация)

K-means кластеры

Всего клиентов 1000

Кластер 0 (Молодые профессионалы) - 300(30%)

- Возраст: до 30 лет
- Доход: средний

- Сумма кредита: низкая
- Высокая активность

Кластер 1 (Семейные клиенты) - 250(25%)

- Возраст: 35–45 лет
- Доход: выше среднего
- Крупные кредиты
- Стабильная активность

Кластер 2 (Премиум-клиенты) - 250(25%)

- Возраст: около 40 лет
- Доход: высокий
- Кредиты на большие суммы
- Часто мужчины

Кластер 3 (Пассивные клиенты) - 150(15%)

- Возраст: старше среднего
- Доход: низкий
- Мало транзакций

Кластер 4 (Потенциальные убывающие) - 50(5%)

- Возраст: около пенсионного возраста
- Доход: низкий
- Почти не пользуются услугами

Отправка результатов:

- Сдача дз через яндекс форму в виде ссылки на ваш github, где будут все файлы и код(python).
- github должен быть открыт и код должен быть рабочий, без ошибок
- Назвать репозитория по шаблону (HW_1_2025-<Имя>_<Фамилия>).
- Ссылка на Яндекс форму https://forms.yandex.ru/u/68eece24505690c23425594c

Дополнительный материал для ДЗ:

- Пример исследования данных с построением сегментаций и выделанию профилений разобранного на семинаре https://colab.research.google.com/drive/15nzMVBa5twjAkAVCf854wkSA9WhanaW5?usp=sharing;
- Пример описания исследования в формате презентации -шпаргалка по домашней работе представлена в папке с вариантами https://disk.yandex.ru/d/i0z880-cKcTSzQ;

- Ссылка на видео описания решения домашней работы 1 (Добавлено будет 16.10.2025 вечером);

GUI:

- https://desktop.github.com/
- https://git-scm.com/
- https://tortoisegit.org/

Материалы :

- - https://git-scm.com/book/ru/v2
- http://www-cs-students.stanford.edu/~blynn/gitmagic/intl/ru/
- https://guides.github.com/activities/hello-world/

Интерактивные туры: https://githowto.com/ru

- https://learngitbranching.js.org/?locale=ru_RU

Полезные ссылки по каждой теме:

- Что база транзакций может рассказать о здоровье вашего бизнеса?

В современном бизнесе данные играют ключевую роль. Одним из важнейших источников информации являются таблицы транзакций. Давайте разберемся, что они могут рассказать о вашем бизнесе и как использовать эту информацию для его улучшения. https://www.youtube.com/watch?v=AFkJbhfXyL8

- Кластеризация: алгоритмы k-means и c-means, https://habr.com/ru/post/67078/
- Clustering performance comparison using K-means and expectation maximization algorithms, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433949/
- EM — масштабируемый алгоритм кластеризации, https://loginom.ru/blog/em
- Ассоциативный анализ рыночной корзины торговой сети, Д. Б. Болдырев, Л. В. Липинский, https://elibrary.ru/item.asp?id=41824659
- How To Perform Market Basket Analysis in Python, https://medium.com/@jihargifari/how-to-perform-market-basket-analysis-in-python-bd00b745b106
- Кластерный анализ, Д.Макров, https://www.dmitrymakarov.ru/intro/clustering-16/
- EM-кластеризация, https://neerc.ifmo.ru/wiki/index.php?title=EM-%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC
- Кластеризация https://academy.yandex.ru/handbook/ml/article/klasterizaciya

- Как увеличить количество повторных заказов в 4,5 раза с помощью RFM-сегментации: кейс омниканального ритейлера «Техносила» – Retail Rocket https://retailrocket.ru/blog/cases/rfm-segmentatsiya-keys-tehnosila/
- Как использовать RFM-анализ в маркетинге https://www.unisender.com/ru/blog/idei/rfm-analiz/
- Using K-means to segment customers based on RFM Variables by Jasneek Chugh Web Mining Medium https://medium.com/web-mining-is688-spring-2021/using-k-means-to-segment-customers-based-on-rfm-variables-9d4d683688c8
- Exploring Customers Segmentation with RFM Analysis and K-Means Clustering with Python. | by Hshan.T | Nov, 2020 | Medium | The Startup https://medium.com/swlh/exploring-customers-segmentation-with-rfm-analysis-and-k-means-clustering-93aa4c79f7a7
- D этой статье есть комбинация и сегментации без учителя и RFM https://habr.com/ru/company/mindbox/blog/423463/
- Портрет клиента

Создание портрета клиента помогает лучше понять целевую аудиторию и адаптировать предложения под ее нужды. Это не миф, а важный инструмент в арсенале аналитика. https://mindbox.ru/journal/cases/12storeez-pryamye-kommunikacii/

- Тут есть много ссылок на полезные ресурсы по построению разных моделей, я его пополняю https://docs.google.com/document/d/13bhjebINH03rGBE-n9n9c4doI22xKrBY7cuQ-GAS3Yw/edit?usp=drivesdk