Задание: Анализ пользовательского поведения в KION[1]

Цель: Исследовать данные о поведении пользователей онлайн-кинотеатра KION, используя статистические методы.
Формат сдачи: отчет в формате html
Дата сдачи: работы принимаются в moodle до 30 марта 23:59
Формат работы: работу можно выполнять как индивидуально, так и в группе до 3 человек. Если вы решаете выполнять работу в группе, в отчете укажите, кто какую роль выполнял. В отчете перечислите ФИО всех членов команды. Отчет загружает ОДИН участник команды.

Данные и их описание

Данные для анализа можно скачать по ссылке.

Описание данных: Датасет включает в себя информацию о взаимодействии пользователей с контентом, демографическую информация о пользователях и мета-информацию о фильмах. Данные собраны на основе анализа пользователей сервиса в период с 13 марта 2021 года по 22 августа 2022 года.

Описание переменных:

user_id — ID пользователя

age — возрастная группа пользователя

- 18_24 — от 18 до 24 лет включительно
- 25_34 — от 25 до 34 лет включительно
- 35_44 — от 35 до 44 лет включительно
- 45_54 — от 45 до 54 лет включительно
- 55_64 — от 55 до 64 лет включительно
- 65_inf — от 65 и старше

sex — пол пользователя

- М — мужчина
- Ж — женщина

income — доход пользователя

[1] KION (КИОН) — российская мультимедийная онлайн-платформа, созданная компанией МТС. Начала работу 20 апреля 2021 года. Kion позволяет смотреть ТВ, сериалы и фильмы на различных устройствах: смартфоне, планшете, компьютере, на Smart TV и ТВ-приставках.

- income_0_20 - доход пользователя от 0 до 20000 р.
- income_20_40 - доход пользователя от 20000 до 40000 р.
- income_40_60 - доход пользователя от 40000 до 60000 р.
- income_60_90 - доход пользователя от 60000 до 90000 р.
- income_90_150 - доход пользователя от 90000 до 150000 р.
- income_150_inf - доход пользователя более 150000 р.

kids_flg — флаг «наличие ребенка»

item_id — ID контента

content_type — Тип контента (фильм, сериал)

title — Название на русском

title_orig — Название оригинальное

genres — Жанры из источника (онлайн-кинотеатры)

countries — страны

for_kids — флаг «контент для детей»

age_rating — возрастной рейтинг

studios — студии

directors — директора

actors — актеры

keywords — ключевые слова

description — описание

valid_from_dttm — дата, с которой контент доступен на KION

rating_kp — рейтинг на Кинопоиске

last_watch_dt — Дата последнего просмотра

total_dur — Общая продолжительность всех просмотров данного контента в секундах

Описание заданий

Ваша задача: познакомить читателя с данными и сделать выводы о пользователях онлайн-кинотеатра КИОН.

Для анализа используйте переменные: **age, income, total_dur, rating_kp, sex.** Проанализируйте каждую переменную, которую будете использовать для выполнения заданий.

Задание №1: Описание данных

Познакомьте читателя с данными:

- Какие данные у вас есть? Какие переменные у вас есть? Какой тип переменных?
- Сколько уникальных пользователей, фильмов и взаимодействий содержится в датасете?

Проведите **описательные статистики** для ключевых переменных (age, income, total_dur, rating_kp, sex).

- Используйте меры центральной тенденции (указывайте, почему выбрали ту или иную меру). Опишите результаты.
- Покажите, как распределены данные (для каждой переменной назовите тип распределения). Используйте графики и меры разброса данных. Какие закономерности видны? Опишите результаты.

Задание №2: Есть ли разница в продолжительности просмотра фильма в зависимости от дохода?

- Определите нулевую и альтернативную гипотезы для ответа на вопрос.
- Укажите, какой тест вы будете использовать для того, чтобы ответить на вопрос и объясните, почему вы его выбрали. Проверьте допущения / предположения теста.
- Выполните необходимые приготовления для теста и проведите тест.
- Опишите результаты и ответьте на вопрос: есть ли разница **в продолжительности просмотра фильма в зависимости от дохода? Пользователи с какой категорией дохода в среднем дольше смотрят фильм? Пользователи с какой категорией дохода в среднем меньше смотрят фильм?** Значимы ли результаты?
- При необходимости визуализируйте результаты.

Задание №3: **Различается ли средняя продолжительность просмотров фильмов между мужчинами и женщинами?**

- Определите нулевую и альтернативную гипотезы для ответа на вопрос.
- Укажите, какой тест вы будете использовать для того, чтобы ответить на вопрос и объясните, почему вы его выбрали. Проверьте допущения / предположения теста.
- Выполните необходимые приготовления для теста и проведите тест.
- Опишите результаты и ответьте на вопрос: кто в среднем смотрит фильмы дольше - мужчины или женщины? Значимы ли результаты?
- Визуализируйте результаты.

Задание №4: Связан ли рейтинг фильма на Кинопоиске и продолжительность просмотра фильмов?

- Определите нулевую и альтернативную гипотезы для ответа на вопрос.
- Укажите, какой тест вы будете использовать для того, чтобы ответить на вопрос и объясните, почему вы его выбрали. Проверьте допущения / предположения теста.
- Выполните необходимые приготовления для теста и проведите тест.
- Опишите результаты и ответьте на вопрос: связан ли рейтинг фильма на Кинопоиске и продолжительность просмотра фильмов. Значимы ли результаты?
- Визуализируйте результаты.

Задание №5: Связан ли пол пользователя и тип контента, который он просматривает?

- Определите нулевую и альтернативную гипотезы для ответа на вопрос.
- Укажите, какой тест вы будете использовать для того, чтобы ответить на вопрос и объясните, почему вы его выбрали. Проверьте допущения / предположения теста.
- Выполните необходимые приготовления для теста и проведите тест.
- Опишите результаты и ответьте на вопрос: связан ли пол пользователя и тип контента, который он просматривает? Какая группа пользователей больше или меньше предпочитает тот или иной контент? Кто больше смотрит сериалы, а кто кино? Приведите конкретные значения. Значимы ли результаты?

- Визуализируйте результаты.

Задание №6: Отличаются ли предпочтения пользователей в типе просматриваемого контента в зависимости от уровня дохода?

- Определите нулевую и альтернативную гипотезы для ответа на вопрос.
- Укажите, какой тест вы будете использовать для того, чтобы ответить на вопрос и почему. Проверьте допущения / предположения теста.
- Выполните необходимые приготовления для теста и проведите тест.
- Опишите результаты и ответьте на вопрос: отличаются ли предпочтения пользователей в типе просматриваемого контента в зависимости от уровня дохода? Какая группа пользователей больше или меньше предпочитает тот или иной контент? Кто больше смотрит сериалы, а кто кино? Приведите конкретные значения. Значимы ли результаты?
- Визуализируйте результаты.

❗Вы должны не только провести тест, но и интерпретировать его результаты, приводя конкретные значения p-value, эффектов и делать осмысленные выводы

Критерии оценивания:

Задание оценивается в диапазоне от **0 до 10 баллов**. Все члены команды получают одинаковую оценку, если задачи внутри команды были распределены **равномерно**.

<u>Что оценивается:</u>

Корректность выбранного статистического теста и аргументация:

- обоснование выбора теста с учетом типа данных и гипотезы
- отражение в отчете логики принятия решения.

Интерпретация каждого результата

- четкое объяснение результатов теста (включая p-value, статистическую значимость, направление и силу эффекта)
- выводы формулируются понятно, без двусмысленности.

Качество оформления отчета

- логичная структура и последовательность изложения.

- удобочитаемость (аккуратность кода, корректное форматирование).
- графики и таблицы оформлены так, чтобы их было легко интерпретировать и понять (подпись осей, заголовков и пр.).

Для оформления отчета используйте: code_folding: show для того, чтобы можно было посмотреть ваш код и скрыть его при необходимости.

Используйте какую-нибудь тему для того, чтобы ваш отчет был оформлен в едином стиле.

Убирайте предупреждения в чанке, используя warning = FALSE.

За что может быть снижена оценка:

Недостаточная аргументация
- Например, в задании просят вычислить **моду**, а студент просто выводит результат, не объясняя его значение.

Использование ИИ без адаптации
- Если интерпретация результатов явно сгенерирована ChatGPT или другим ИИ и не переработана студентами.

Неясность или поверхностность выводов
- Например, если результаты описаны абстрактно, без конкретных числовых значений или без связи с контекстом задачи.

Просроченная сдача
- Если работа не сдана в срок, выставляется **0 баллов** независимо от ее качества.