Сборный проект

Теперь самое время проверить знания и решить аналитический кейс. Выполнять работу вы будете самостоятельно.
Когда закончите, отправьте работу на проверку ревьюеру. В течение суток вы получите комментарии. Их нужно учесть: доработать проект и вернуть ревьюеру обновлённый вариант. Возможно, вы будете дорабатывать кейс по комментариям несколько раз. Не переживайте, это нормально.
Проект завершён, когда одобрены все доработки.

Описание проекта

Заказчик этого исследования — Министерство культуры Российской Федерации.
Вам нужно изучить рынок российского кинопроката и выявить текущие тренды. Уделите внимание фильмам, которые получили государственную поддержку. Попробуйте ответить на вопрос, насколько такие фильмы интересны зрителю.
Вы будете работать с данными, опубликованными на портале открытых данных Министерства культуры. Набор данных содержит информацию о прокатных удостоверениях, сборах и государственной поддержке фильмов, а также информацию с сайта КиноПоиск.

Инструкция по выполнению

Шаг 1. Откройте файлы с данными и объедините их в один датафрейм
Объедините данные таким образом, чтобы все объекты из датасета `mkrf_movies` обязательно вошли в получившийся датафрейм.
Пути к файлам:
`/datasets/mkrf_movies.csv` — данные о прокатных удостоверениях.
Скачать датасет
`/datasets/mkrf_shows.csv` — данные о прокате в российских кинотеатрах.
Скачать датасет
Шаг 2. Предобработка данных

- Проверьте типы данных в датафрейме и преобразуйте там, где это необходимо.
- Изучите пропуски в датафрейме. Объясните, почему заполнили пропуски определённым образом или почему не стали это делать.
- Проверьте, есть ли в данных дубликаты. Опишите причины, которые могли повлиять на появление дублей.
- Изучите столбцы, которые содержат категориальные значения:
    - Посмотрите, какая общая проблема встречается почти во всех категориальных столбцах;
    - Исправьте проблемные значения в поле `type`.
- Изучите столбцы, которые хранят количественные значения. Проверьте, обнаружились ли в таких столбцах подозрительные данные. Как с такими данными лучше поступить?
- Добавьте новые столбцы:
    - Создайте столбец с информацией о годе проката. Выделите год из даты премьеры фильма;
    - Создайте два столбца: с именем и фамилией главного режиссёра и основным жанром фильма. В столбцы войдут первые значения из списка режиссёров и жанров соответственно;

- Посчитайте, какую долю от общего бюджета фильма составляет государственная поддержка.

Шаг 3. Проведите исследовательский анализ данных

- Посмотрите, сколько фильмов выходило в прокат каждый год. Обратите внимание, что данные о прокате в кинотеатрах известны не для всех фильмов. Посчитайте, какую долю составляют фильмы с указанной информацией о прокате в кинотеатрах.
- Изучите, как менялась динамика проката по годам. В каком году сумма сборов была минимальной? А максимальной?
- С помощью сводной таблицы посчитайте среднюю и медианную сумму сборов для каждого года.
- Определите, влияет ли возрастное ограничение аудитории («6+», «12+», «16+», «18+» и т. д.) на сборы фильма в прокате в период с 2015 по 2019 год? Фильмы с каким возрастным ограничением собрали больше всего денег в прокате? Меняется ли картина в зависимости от года? Если да, предположите, с чем это может быть связано.

Шаг 4. Исследуйте фильмы, которые получили государственную поддержку
На этом этапе нет конкретных инструкций и заданий — поищите интересные закономерности в данных. Посмотрите, сколько выделяют средств на поддержку кино. Проверьте, хорошо ли окупаются такие фильмы, какой у них рейтинг.
Шаг 5. Напишите общий вывод

Оформление

Выполните задание в Jupyter Notebook. Заполните программный код в ячейках типа *code*, текстовые пояснения — в ячейках типа *markdown*. Используйте форматирование и заголовки.

Описание данных

Таблица `mkrf_movies` содержит информацию из реестра прокатных удостоверений. У одного фильма может быть несколько прокатных удостоверений.

- `title` — название фильма;
- `puNumber` — номер прокатного удостоверения;
- `show_start_date` — дата премьеры фильма;
- `type` — тип фильма;
- `film_studio` — студия-производитель;
- `production_country` — страна-производитель;
- `director` — режиссёр;
- `producer` — продюсер;
- `age_restriction` — возрастная категория;
- `refundable_support` — объём возвратных средств государственной поддержки;
- `nonrefundable_support` — объём невозвратных средств государственной поддержки;
- `financing_source` — источник государственного финансирования;
- `budget` — общий бюджет фильма;
- `ratings` — рейтинг фильма на КиноПоиске;
- `genres` — жанр фильма.

Обратите внимание, что столбец `budget` уже включает в себя полный объём государственной

поддержки. Данные в этом столбце указаны только для тех фильмов, которые получили государственную поддержку.
Таблица `mkrf_shows` содержит сведения о показах фильмов в российских кинотеатрах.

- `puNumber` — номер прокатного удостоверения;
- `box_office` — сборы в рублях.

Как будут проверять мой проект

Ваш проект будут оценивать по конкретным критериям. Прежде чем решать кейс, внимательно изучите их.
На что обращают внимание при проверке проектов:

- как вы описываете найденные в данных проблемы;
- какие методы замены типов данных, обработки пропусков и дубликатов применяете и как обосновываете принятое решение;
- автоматизируете ли вы однотипные действия;
- выводите ли финальные данные в сводных таблицах;
- какие методы построения графиков применяете;
- соблюдаете ли структуру проекта и поддерживаете аккуратность кода;
- какие выводы делаете;
- оставляете ли комментарии к шагам.

Всё необходимое для того, чтобы выполнить проект, есть в темах, которые вы прошли.
Успехов!