

**Дипломная работа по курсу “Аналитик данных. Старт в профессии”**

**Студент: Самойлов Павел Александрович**

**Набор: с 01 ноября 2021 года в рамках проекта «Цифровые профессии»**

**Дипломная работа по курсу «Аналитик данных. Старт в профессии»**

|  |  |
| --- | --- |
| **Задание 1. Дать ответы на поставленные вопросы по заданному набору данных “**MovieLens**”** | |
| 0. Подготовить данные для анализа | Необходимо несколько слов сказать о подготовке данных. Данные из трех таблиц задания были выгружены в Excel. Последующая их обработка производилась уже с помощью этого инструмента. В табличный формат данные были переведены посредством опции разделения по столбцам, возникала проблема с тем, что при выгрузке по некоторым строкам данные сохранялись не в одну ячейку, а в несколько, но эту проблему удалось легко решить использованием функции = СЦЕПИТЬ (). Для удобства последующего использования к таблицам с данными была применена опция создания «умной» таблицы (Ctrl + T), это позволило в дальнейшем избежать сложностей с растягиванием формул по значительному количеству строк. Следующий этап — это перевод столбца с датой из формата UNIX в нормальную дату (сделал через формулу исходя из того что 01.01.1970 начинает отсчет не с 0 секунд, а с 25569, в одном дне 86400 секунд, ну а дальше сложение с базовой датой и получаем нормальный формат, выставил США «число – месяц – год», потому что в данных уже есть сведения в таком формате). В итоге данные были сформированы в 3 листа (u.item, u.user, u.data) в файл «ДАННЫЕ ПО ФИЛЬМАМ корректированный». Указанной первоначальной подготовки данных было достаточно для их выгрузки и анализа с использованием Power BI (правда, чтобы соединить все три таблицы по полю item id в табличке со списком фильмов пришлось переименовать столбец movie id в столбец item id. В результате данные подтянулись автоматически):    Но, к сожалению, у инструмента Power BI есть сложности с представлением и выгрузкой визуальной информации в другие документы (дополнительные программы и рабочая учетная запись), поэтому дополнительно была сформирована сводная таблица в формате Excel (файл – «СВОДНАЯ ТАБЛИЦА по оценкам фильмов»), источником для нее послужила таблица с оценками фильмов в которую с использованием функции ВПР были подтянуты данные по фильмам и по людям которые их оценивали – то есть в каждой строчке такой таблицы были собраны все данные из трех таблиц («умышленная» денормализация). В этой же книге добавлены листы для некоторых аналитических инструментов и диаграмм. Для регрессионного анализа был сформирован файл «date1.txt» в который были выгружены 100 000 строк, содержащих по столбцам 19 жанров и рейтинговую оценку (от 1 до 5), поставленную пользователем. Все файлы приложены. |
| 1. Определите динамику оценок по месяцам | Анализ динамики оценок по месяцам показал минимальные значения в сентябре 1997 года т.е. в месяц начала сбора данных (6863 оценок). Максимальное количество оценкок зафиксировано в ноябре 1997 года (24118). В дальнейшем показатели фиксировались на уровне 10 000 с явной тенденцией на снижение. В апреле 1998 года количество оценок составило 8889 (возможно это связано с утратой интереса к эксперименту, а этот сбор данных был именно университетстским экспериментом т.е не коммерческим сбором данных. Поэтому я бы с осторожностью относился с прогнозу который изображен на графике (небольшой рост), на мой взгляд сезонности в этих данных нет. |
| 2. Определите растет ли средняя оценка или падает? | Со средней оценкой получилось интереснее. В начале эксперимента в сентябре 1997 года и до конца 1997 года отмечаются стабильные средние значения в районе 3,5. В январе 1998 года мы видим резкое снижение до 3,4 и дальнейший рост уже к марту 1998 года. Общий наблюдаемый тренд указывает на снижение средней оценки, прогноз также говорит о снижении в будущем.      Здесь я тоже не вижу сезонности. Что касается причин то анализ по сводной таблице показал, что причина такого низкого рейтинга в январе 1998 года состоит в том, что к эксперименту подключились несколько человек которые поставили очень много негативных оценок:    «Лидером» здесь стала 22 летняя медсестра, которая умудрилась поставить 485 оценок на 1 балл. (думаю, что такие оценки можно считать выбросами и исключать, но я для чистоты эксперимента оставил). |
| 3. Определите как много уникальных пользователей ставят оценки | Ответ на вопрос есть в описании датасета:  «*Этот набор данных состоит из:*  *\* 100 000 оценок (1–5) от 943 пользователей на 1682 фильмах.*  *\* Каждый пользователь оценил не менее 20 фильмов.*  *\* Простая демографическая информация для пользователей (возраст, пол, род занятий, почтовый индекс)»*  Но думаю, стоит это проверить, самый простой способ — это посчитать количество уникальных элементов в столбце user id, я сделал это несколькими способами, но наиболее изящный мне показался Power BI, в этой программе соответствующая опция есть прямо на виду при выборе типа агрегации данных: |
| 4. Составьте топ фильмов по количеству оценок | По количеству оценок, а данных лидирует фильм «Звездные Войны (1977 года). Он набрал 583 оценки, но думаю существенный интерес представляют не количество оценок, а рейтинги. В это контексте в сводной таблице я сопоставил среднюю оценку, количество оценок и стандартное отклонение. Как результат могу смело утверждать, указанный мной фильм «Звездные войны» является лучшим по оценкам американских зрителей в 1997–1998 годах. Более того стандартное отклонение у этого фильма одно из самых низких – 0,88 это говорит о низко разбросе оценок, то есть зрители в основном ставят ему стабильно веские оценки.  В полученной выборке я обратил внимание на фильмы с высокой средней оценкой и высоким количеством оценок, что позволило сформировать своеобразный «ТОП лист».  Ниже на странице приведены: ТОП по лист по количеству оценок, ТОП выжимка лучших фильмов:      А вот эта таблица топ, по средней оценке, но доверять этой оценке нельзя, она сделана по 1-2 оценкам. |
| 5. Изучите клиентскую базу и выведите основные группы клиентов (возраст, профессия. | Очень интересное задание. Первое что я сделал сформировал таблицу по профессиям в характеристики были выбраны: количество уникальных лиц, количество проставленных ими отметок, средний рейтинг, медианный рейтинг и стандартное отклонение. Отдельным полем был добавлен коэффициент активности, то есть количество оценок, деленное на количество уникальных лиц их сделавших. По этому ключу и была сделана сортировка по убыванию:    Как мы видим, наиболее активно оценивали фильмы медсестры, но мы помним, что во многом этому способствовала наша 22 летняя медсестра под номером 405, наверное, по ее же вине у медсестер самые негативные оценки. Наименьшую активность проявили домохозяйки, всемером они поставили 299 оценок. В центре внимания студенты – их больше всего приняло участие в исследовании – 196, они же проставили наибольшее количество голосов – 21957, у них относительно высокий коэффициент активности, они не токсичны их средние оценки очень близки с общей средней. Стоит обратить внимание так же на категорию «другая сфера занятости» очевидно – что того места, где люди работают просто, не было в опроснике, а поле others не предусматривало самостоятельного заполнения места работы. Не стоит упускать из виду активно голосующих программистов и технологов, учителей и администраторов.  По гендерному признаку. Всего в оценивании приняло участие 670 мужчин и 273 женщины, средний рейтинг у полов идентичный, мужчины голосовали чуть-чуть активнее |
| **Задание 2: проанализировать данные и на их основе разработать рекомендации по совершенствованию бизнеса (онлайн – кинотеатра).** | |
| Опишите основные бизнес-отчеты (2–3 штуки), которые мы хотим видеть по бизнесу | Решение |
| Опишите основные имеющиеся данные и источники их поступления | Набор данных MovieLens были собраны исследовательским проектом GroupLens в Университете Миннесоты.    Этот набор данных состоит из:  \* 100 000 оценок (1–5) от 943 пользователей на 1682 фильмах.  \* Каждый пользователь оценил не менее 20 фильмов.  \* Простая демографическая информация для пользователей (возраст, пол, род занятий, почтовый индекс)  Данные были собраны через веб-сайт MovieLens в семимесячный период с 19 сентября 1997 г. по 22 апреля 1998 г. Эти данные очищены, то есть пользователи, которые имели менее 20 оценок или не имели полной демографической информация была удалена из этого набора данных.  **u.data** - Полный набор данных u, 100000 оценок 943 пользователями по 1682 элементам. Каждый пользователь оценил не менее 20 фильмов. Пользователи и предметы пронумерованы последовательно от 1. Данные перемешаны случайным образом. Это список разделенных табуляцией показателей:  идентификатор пользователя | идентификатор товара | рейтинг | отметка времени (Временные метки - unix-секунды с 01.01.1970 по всемирному координированному времени)  **u.item** - Информация о фильмах; это список разделенный табуляцией:  идентификатор фильма | название фильма | дата выпуска | дата выпуска видео |IMDb URL | неизвестно | Действие | Приключение | Анимация |Детская | Комедия | Преступление | Документальный фильм | Драма | Фэнтези |Фильм-нуар | Ужас | Музыкальный | Тайна | Романтика | Научная фантастика |Триллер | Война | Западный |  Последние 19 полей — это жанры, 1 – фильм относится к этому жанру, 0 означает, что фильм не относится к этому жанру; фильмы могут быть сразу несколько жанров. Идентификаторы фильмов используются в наборе данных u.data.  **u.user** - Демографические данные о пользователях; это вкладка разделенный список:  идентификатор пользователя | возраст | пол | занятие | индекс  Идентификаторы пользователей используются в наборе данных u.data, то есть связаны с ним по внешнему ключу. |
| Придумайте Data-проект, который должен улучшить показатели бизнеса | Перспектива для дальнейшей аналитики:  Эти данные очищены – пользователи, которые имели менее 20 оценок или не имели полной демографической информация была удалена из этого набора данных. |