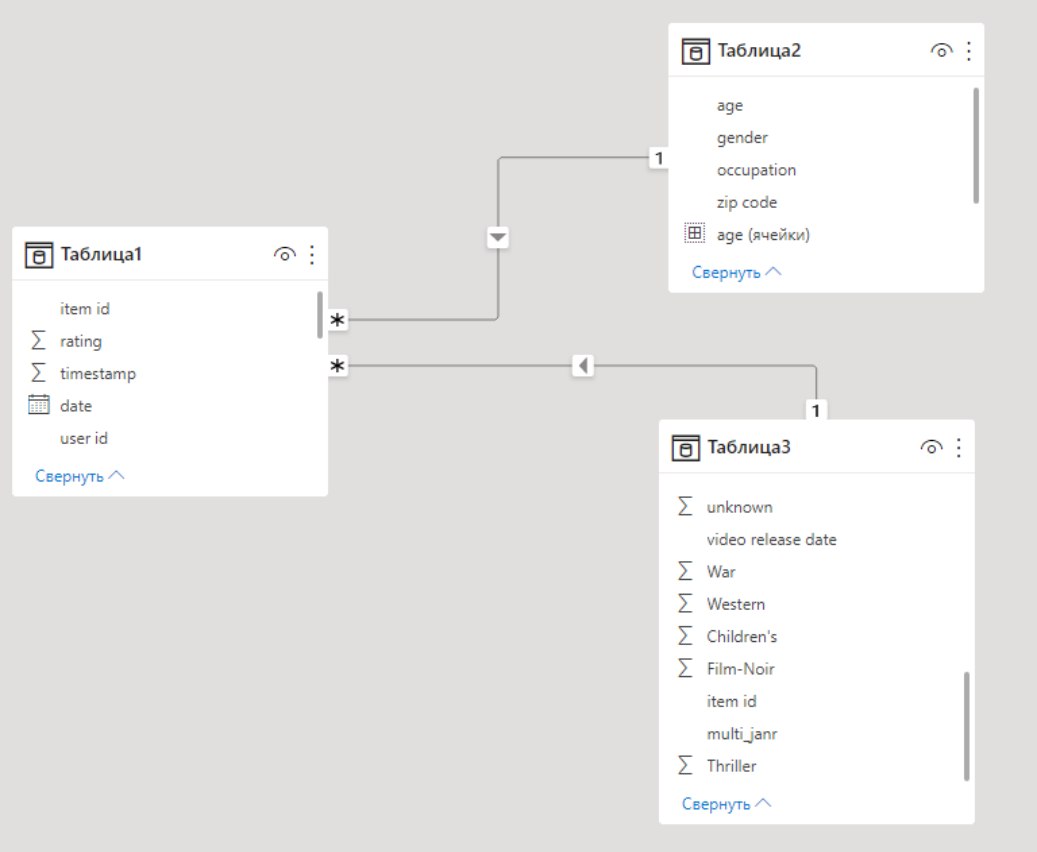
Дипломная работа по курсу "Аналитик данных. Старт в профессии"

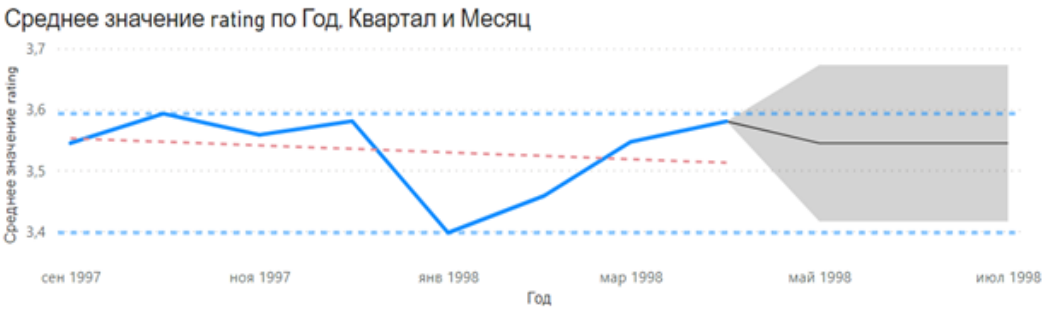
Студент: Самойлов Павел Александрович

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

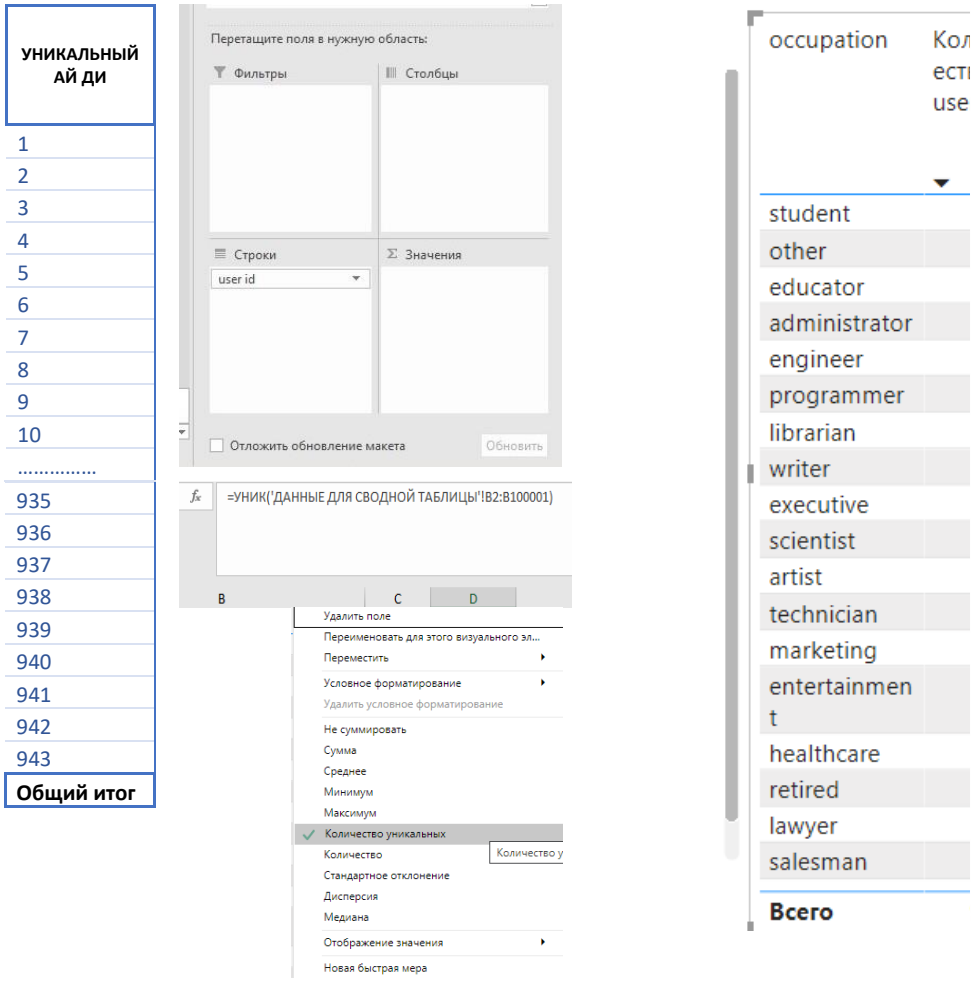
Дипломная работа по курсу «Аналитик данных. Старт в профессии»

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.  Год | 1997 | | | | 1998 | | | | Месяц | Сентябрь | Октябрь | Ноябрь | Декабрь | Январь | Февраль | Март | Апрель Количество rating | 6863 | 10313 | 24118 | 11650 | 14154 | 11016 | 12997 | 8889
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.  Год | 1997 | | | | 1998 | | | | Месяц | Сентябрь | Октябрь | Ноябрь | Декабрь | Январь | Февраль | Март | Апрель Среднее значение rating | 3,54 | 3,59 | 3,56 | 3,58 | 3,4 | 3,46 | 3,55 | 3,58

	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: date (несколь элементов) Сумма по полю СЧЕТ УНИКАЛЬНЫЙ АЙ ДИ | Названия 1 | 2 | 3 | 4 | 5 | Общий итог 405 | 485 | 73 | 63 | 48 | 68 | 737 537 | 52 | 103 | 204 | 121 | 10 | 490 846 | 10 | 46 | 89 | 154 | 106 | 405 201 | 34 | 82 | 132 | 114 | 24 | 386 435 | 18 | 63 | 126 | 111 | 55 | 373 561 | 37 | 45 | 154 | 100 | 16 | 352 727 | 15 | 92 | 109 | 78 | 28 | 322 «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:

НАИМЕНОВАНИЯ ФИЛЬМОВ	Среднее по полю rating	Количество по полю rating	Стандартное отклонение по полю rating
Star Wars (1977)	4,4	583	0,881341012
Contact (1997)	3,8	509	0,994427025
Fargo (1996)	4,2	508	0,975755983
Return of the Jedi (1983)	4,0	507	0,923954854
Liar Liar (1997)	3,2	485	1,098544042
English PatientThe (1996)	3,7	481	1,169400816
Scream (1996)	3,4	478	1,113910362
Toy Story (1995)	3,9	452	0,927896701
Air Force One (1997)	3,6	431	0,998071847
Independence Day (ID4) (1996)	3,4	429	1,11658384
Raiders of the Lost Ark (1981)	4,3	420	0,89181858
GodfatherThe (1972)	4,3	413	0,934576797
Pulp Fiction (1994)	4,1	394	1,150879817
Twelve Monkeys (1995)	3,8	392	0,982036979
Silence of the LambsThe (1991)	4,3	390	0,836596997
Jerry Maguire (1996)	3,7	384	0,937907875
Chasing Amy (1997)	3,8	379	1,053065641
RockThe (1996)	3,7	378	0,980448277
Empire Strikes BackThe (1980)	4,2	367	0,922803143
Star Trek: First Contact (1996)	3,7	365	0,928596324
Titanic (1997)	4,2	350	0,946821816

РАНЖИР	НАИМЕНОВАНИЯ ФИЛЬМОВ	Среднее по полю rating	Количество по полю rating	Стандартное отклонение по полю rating
1	Star Wars (1977)	4,4	583	0,881341012
2	Schindler's List (1993)	4,5	298	0,829109354
3	Raiders of the Lost Ark (1981)	4,3	420	0,89181858
4	GodfatherThe (1972)	4,3	413	0,934576797
5	Fargo (1996)	4,2	508	0,975755983
6	Silence of the LambsThe (1991)	4,3	390	0,836596997
7	Pulp Fiction (1994)	4,1	394	1,150879817
8	Empire Strikes BackThe (1980)	4,2	367	0,922803143
9	Titanic (1997)	4,2	350	0,946821816
10	Princess BrideThe (1987)	4,2	324	0,998948344
11	Monty Python and the Holy Grail (1974)	4,1	316	0,994595603
12	Return of the Jedi (1983)	4,0	507	0,923954854
13	Toy Story (1995)	3,9	452	0,927896701

А вот эта таблица топ, по средней оценке, но доверять этой оценке нельзя, она сделана по 1–2 оценкам.

НАИМЕНОВАНИЯ ФИЛЬМОВ	Среднее по полю rating	Количество по полю rating
Someone Else's America (1995)	5	1
Santa with Muscles (1996)	5	2
They Made Me a Criminal (1939)	5	1
Saint of Fort WashingtonThe (1993)	5	2
Star Kid (1997)	5	3
Prefontaine (1997)	5	3
Marlene Dietrich: Shadow and Light (1996)	5	1
Great Day in HarlemA (1994)	5	1
Entertaining Angels: The Dorothy Day Story (1996)	5	1
Aiqing wansui (1994)	5	1
Pather Panchali (1955)	4,625	8
Some Mother's Son (1996)	4,5	2
Maya Lin: A Strong Clear Vision (1994)	4,5	4
Everest (1998)	4,5	2
Anna (1996)	4,5	2

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:

occupation	Количество user id	Количество rating	коэффициент активности (количество голосов / на количество проголосовавших)	Среднее значение rating	Медиана rating	Стандартное отклонение для rating
healthcare	16	2804	175	2,9	3	1,27
technician	27	3506	130	3,53	4	1,05
writer	45	5536	123	3,38	3	1,2
engineer	67	8175	122	3,54	4	1,07
programmer	66	7801	118	3,57	4	1,12
entertainment	18	2095	116	3,44	4	1,18
retired	14	1609	115	3,47	4	0,96
lawyer	12	1345	112	3,74	4	1,1
student	196	21957	112	3,52	4	1,13
executive	32	3403	106	3,35	4	1,32
librarian	51	5273	103	3,56	4	1,05
other	105	10663	102	3,55	4	1,11
none	9	901	100	3,78	4	1,08
educator	95	9442	99	3,67	4	1,08
administrator	79	7479	95	3,64	4	1,09
artist	28	2308	82	3,65	4	1,18
doctor	7	540	77	3,69	4	0,96
marketing	26	1950	75	3,49	4	1,08
salesman	12	856	71	3,58	4	1,18
scientist	31	2058	66	3,61	4	1,01
homemaker	7	299	43	3,3	3	1,17

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

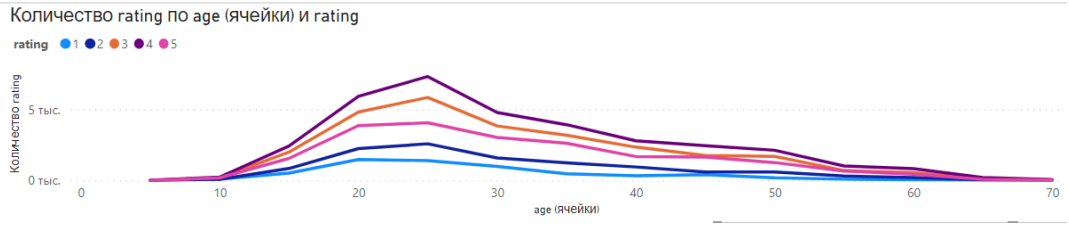
активности, они не токсичны их средние оценки очень близки с общей средней. Стоит обратить внимание так же на категорию «другая сфера занятости» очевидно – что того места, где люди работают просто, не было в опроснике, а поле others не предусматривало самостоятельного заполнения места работы. Не стоит 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

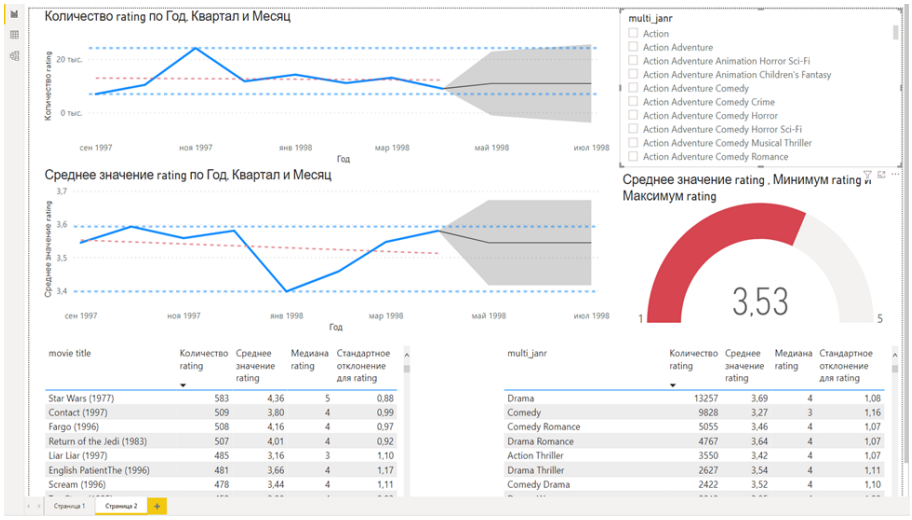
gender	Количество user id	Среднее значение rating	Медиана rating	Стандартное отклонение для rating	Количество rating разделить на Количество gender 3
F	273	3,53	4	1,17	94
M	670	3,53	4	1,11	111

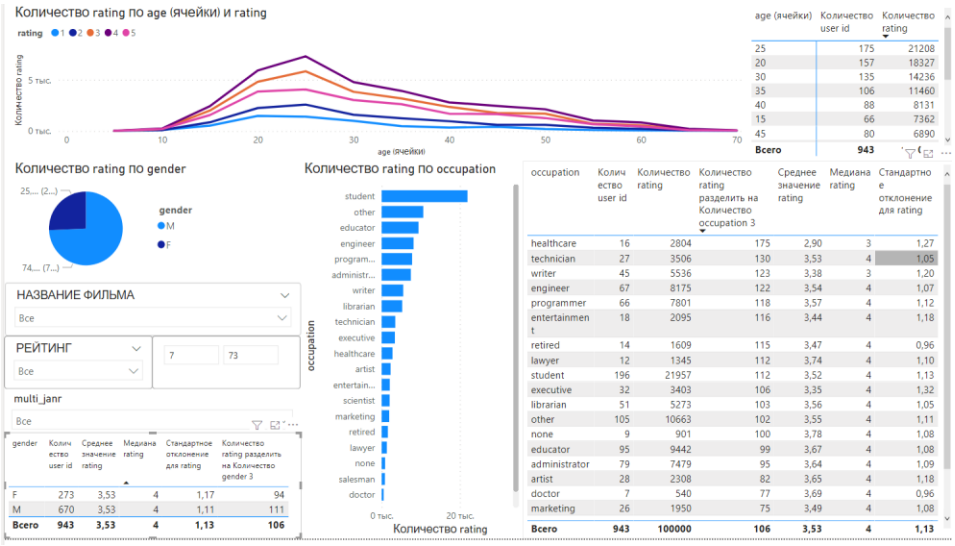
age (группы)	Количество rating	Количество user id
25	21208	175
20	18327	157
30	14236	135
35	11460	106
40	8131	88
15	7362	66
45	6890	80
50	5890	62
55	2814	32
60	2109	19
10	819	10
65	514	8
70	197	4
5	43	1

По возрасту.
для удобства возраст был поделен на группы с интервалом по 5 лет. В итоге мы получили наиболее активную аудиторию от 20 до 35 лет. На это стоит обратить внимание. Они же ставят и наиболее высокие оценки.



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	На сколько я понял в этот блок необходимо включить отчеты, которые были бы полезны бизнесу. Основной интерес на мой взгляд для бизнеса состоит в определении целевой аудитории и фильмов, которые ей нравятся. Посылка следующая – прежде чем оценить фильм человек его покупает, то есть мы понимаем, что человеку нравится, соответственно мы можем разрабатывать рекомендательные системы и поддерживать потребительскую активность зрителей. Конечно, создание рекомендательной системы для меня задача пока очень сложная и далекая, поэтому в рамках этой работы я сделал лишь 2 листа отчетов, которые помогают определить целевую аудиторию, и фильмы, которые пользуются популярностью. Отчасти я отметил некоторые из этих показателей в отчете по Заданию № 1. Дополнительно мне показалась интересной мысль об исследовании мультижанровости фильмов, то есть в одну ячейку мной были сведены значения всех жанров, относящихся к этому фильму. Это позволит выявить наиболее привлекательные сочетания жанров. Отчеты сделаны в Power BI, мне чрезвычайно понравился этот инструмент, однако, есть у него проблемы в выгрузку отчетов. Точнее не у инструмента, а у меня – у меня нет рабочей учетной записи микрософт, а все попытки ее создать приводили к ошибкам (судя по поисковым запросам в Google не у меня одного). Поэтому в качестве решения этого задания я направляю файл power BI вот с этими отчетами: 

	 Я вполне осознаю, что отчеты не очень красочные и содержат множество таблиц, но мне пока удобнее работать с таблицами, то есть это отчеты как инструмент анализа, а не презентация товаров и проблем перед советом акционеров, но прошу оценить то, что у меня получилось (если необходимо все-таки выгрузка буду осваивать Tableau – хотя мне он нравится меньше). Файл «отчеты по кинотеатру» в приложении к письму.
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.

Определенные аналитики были описаны мной в предыдущих разделах, в этом блоке мне бы хотелось привести анализ по жанрам и мультижанрам.

Для начала обратите внимание на мультикорреляцию (рисунок ниже), конечно наибольший интерес представляет первый столбец – то есть корреляционная зависимость жанра с рейтингом, но об этом чуть позже.

	rating	unknown	Action	Adventure	Animation	Children's	Comedy	Crime	ocumentai	Drama	Fantasy	Film-Noir	Horror	Musical	Mystery	Romance	Sci-Fi	Thriller	War	Western
rating	1																			
unknown	-0,00293	1																		
Action	-0,02585	-0,00586	1																	
Adventure	-0,00934	-0,00399	0,451525	1																
Animation	0,008047	-0,00193	-0,09902	-0,02473	1															
Children's	-0,04364	-0,00278	-0,14484	0,100567	0,555386	1														
Comedy	-0,07865	-0,00652	-0,22328	-0,11301	0,029612	0,082916	1													
Crime	0,02693	-0,00296	0,007478	-0,03007	-0,05724	-0,08233	-0,09099	1												
Documentary	0,011099	-0,00087	-0,05125	-0,0349	-0,0169	-0,02431	-0,05673	-0,02587	1											
Drama	0,114006	-0,00815	-0,26983	-0,2237	-0,15755	-0,12997	-0,34746	0,064043	-0,05778	1										
Fantasy	-0,03272	-0,00117	-0,01388	0,08776	0,026133	0,238081	0,01716	0,006394	-0,01023	-0,02076	1									
Film-Noir	0,046206	-0,00133	-0,07788	-0,05303	-0,02568	-0,03694	-0,08659	0,163712	-0,01161	-0,08284	-0,01555	1								
Horror	-0,05041	-0,00237	-0,00772	-0,05865	-0,02837	-0,06592	-0,07405	-0,01544	-0,02071	-0,15948	-0,02774	-0,03147	1							
Musical	-0,00172	-0,00228	-0,09121	-0,0248	0,417647	0,381293	0,035065	-0,06757	-0,01995	-0,0966	-0,02673	-0,03032	-0,0541	1						
Mystery	0,02263	-0,00235	-0,03281	-0,04368	-0,0455	-0,05502	-0,11135	0,087773	-0,02056	-0,06866	-0,02754	0,232057	0,001824	-0,05371	1					
Romance	0,040107	-0,00492	-0,01875	-0,01852	-0,08544	-0,11932	0,095863	-0,10252	-0,04296	0,013052	-0,01708	-0,05521	-0,07607	-0,01014	-0,05991	1				
Sci-Fi	0,010471	-0,00382	0,323875	0,294811	-0,04473	-0,04221	-0,14562	-0,08692	-0,03338	-0,17435	0,125709	0,016183	0,034251	-0,08153	-0,03078	-0,06331	1			
Thriller	-0,0098	-0,00529	0,249647	-0,04945	-0,07715	-0,14418	-0,29053	0,124021	-0,04624	-0,16336	-0,04728	0,11029	0,069872	-0,11143	0,230351	-0,10605	0,046936	1		
War	0,081815	-0,00322	0,166665	0,087115	-0,0564	-0,08534	-0,11982	-0,09533	-0,00721	0,098864	-0,0377	-0,04277	-0,07632	-0,05553	-0,07577	0,126645	0,167349	-0,10027	1	
Western	0,010184	-0,00137	0,063467	0,010551	-0,02658	-0,03105	0,002417	-0,04068	-0,01201	-0,03325	-0,01609	-0,01825	-0,03257	-0,03138	-0,03234	-0,05239	-0,05249	-0,07272	-0,02369	1

Сначала я бы хотел обратить внимание на сочетание жанров.

Интересно, например, что корреляция между фильмами для детей и мультфильмами – 0,55. Думаю это стоит понимать как то, что фильмы для детей в основной своей массе анимационные. Равно как драма абсолютно не сочетается с комедией (коэффициент 0,34). (Другие интересные сочетания и «сочетания смотрите на рисунке по цветам).

	КОРРЕЛЯЦИЯ С РЕЙТИНГОМ	коэффициенты регрессии полученные с помощью OCTAVE	количество ФИЛЬМОВ С ОТМЕТКОЙ О ПРИНАДЛЕЖНОСТИ К ЖАНРУ
Drama	0,11400597	0.243738	39895
Comedy	-0,078653489	-0.063683	29832
Action	-0,02584682	-0.085037	25589
Thriller	-0,009801672	0.022260	21872
Romance	0,040107392	0.114801	19461
Adventure	-0,009341679	0.081709	13753
Sci-Fi	0,010471415	0.096023	12730
War	0,081814724	0.254046	9398
Crime	0,02692992	0.092779	8055
Children's	-0,043644125	-0.219883	7182
Horror	-0,050412632	-0.132037	5317
Mystery	0,022629541	0.111072	5245
Musical	-0,001716456	0.062822	4954
Animation	0,008046795	0.356888	3605
Western	0,010184017	0.194008	1854
Film-Noir	0,046205969	0.396209	1733
Fantasy	-0,032720866	-0.204077	1352
Documentary	0,011099435	0.254059	758
unknown	-0,002930496	-0.182756	10

Перейдем к рейтингам. Конечно, мульти корреляция не 100 инструмент, но все же мы видим, что у драматических фильмов этот коэффициент самый большой (понятно, что он меньше 0,5, но все же самый большой). Это означает, что, если у фильма есть отметка с отнесением к жанру драма его шанс получить высокие рейтинги самый большой. С этой же позиции внимания заслуживают: военные фильмы и криминальные драмы. Напротив, снять хорошую комедию очень тяжело – и об этом говорит самая низкая корреляция этого жанра с рейтингом (с осторожностью следует готовить к прокату в нашем кинотеатре мультфильмы и ужасы, у них капризная аудитория). Приведенный анализ я попытался подтвердить или опровергнуть путем вычисления коэффициентов регрессии. Вычисления производились посредством нормального уравнения в среде Octave путем загрузки датасета из 19 столбцов, содержащих отметку (1 или 0) с принадлежностью к жанру – обозначены за X, рейтинг по каждой из 100 000 составил Y. Результат: базовое значение составило 3,38, а коэффициенты по каждому жанру приведены в таблице выше. В целом наши выводы подтвердились, и мы видим самые большие коэффициента у драматических фильмов, военных фильмов, криминальных драм и документального кино. Но мы должны понимать, что высокие оценки — это еще не все у каждого жанра – своя аудитория, но это уже другое сложное исследование. Вместе с этим стоит обратить внимание на то, что у нашей библиотеке мало фильмов перспективных направлений таких как документальные и криминальные драмы, над этим стоит поработать.

Придумайте Data-проект, который должен улучшить показатели бизнеса	Наш проект напрямую связан с результатом анализа данных, поэтому в первую очередь об этом, то есть не о бизнесе, а о работе с данными. Что бы получить более интересные результаты нам стоит обратить внимание на то, что 105 наших респондентов относятся к категории others – то есть, по сути, мы не знаем кто эти люди, а они стоят на втором месте после студентов не только по количеству участников, но и по количеству оценок. То есть мы не знаем ничего об 1/10 наших потребителей – этого допускать нельзя и при следующем сборе данных необходимо как минимум предлагать человеку указать свою профессию самому. Да это будет сложно анализировать, но так мы не потеряем данные. То же самое можно сказать и про очистку данных: *«Эти данные очищены – пользователи, которые имели менее 20 оценок или не имели полной демографической информация была удалена из этого набора данных»* То есть мы теряем из виду людей, которые просто проявили меньшую активность в голосовании, но это все те же наши потенциальные потребители, не учитывать их мнение неправильно (но учитывать при этом медсестру № 405). Предлагается при дальнейших исследованиях не игнорировать лиц, проголосовавших менее 20 раз, может им просто порекомендовали не тот фильм, и они ушли к конкурентам, а может вопрос в интерфейсе. Теперь к предложениям по бизнесу. Сложно. Я всю жизнь был связан только с юриспруденцией, в этой плоскости я понимаю и статистику, и процессы. В бизнесе все немного по-другому, а в кинобизнесе наверняка куча особенностей. Поэтому нам нужен эксперт не только по данным, но и по кино. Что бы «крутить данные вместе». Понятно, что привлечение специалиста стоит денег, но их стоит потратить. Далее аудитория: 20–35 лет, студенты, администраторы, программисты, техники, учителя – от наша аудитория. Поэтому для продвижения бизнеса стоит подумать о запуске целевой рекламы на тех ресурсах Интернета, которыми пользуются указанные категории людей, и в местах, которые эти люди посещают (учебные заведения, кафе) да это требует дополнительного анализа, но в развитии бизнеса нет мелочей. Следующий блок жанры. Сюрприз исследования — это

криминальные драмы они очень нравятся зрителю, но их в нашей библиотеке очень мало, это надо исправить. Что это за фильмы? Топ 5 приведен ниже. Обратите внимание на то, что стандартное отклонение у них очень низкое (в сравнении, разумеется), рейтинги высокие и количество голосов не минимальные (которые нужно игнорировать). То есть если человек посмотрел этот фильм, то очень вероятно он поставил ему высокую оценку, значит такой филь можно рекомендовать и другим пользователям. Лично по мне помню восторг от фильма «Спрут», американцы 1997–1998 похоже просто его не смотрели, иначе он тоже был бы в топе.

Film-Noir	1		
НАИМЕНОВАНИЯ ФИЛЬМОВ	**Среднее по полю rating**	**Количество по полю rating**	**Стандартное отклонение по полю rating**
L.A. Confidential (1997)	4,2	297	0,854721434
Blade Runner (1982)	4,1	275	0,905677733
Chinatown (1974)	4,1	147	0,983699028
Maltese FalconThe (1941)	4,2	138	0,758916809
Manchurian CandidateThe (1962)	4,3	131	0,760443052

По жанрам я бы так же порекомендовал продолжить (в том числе с экспертом по кино) накапливать фильмы драматического содержания, в том числе и военные драмы, ну и обратить внимание на космические приключенческие боевики, поскольку эти жанры имеют неплохие показатели по рейтингам в сочетании с количеством просмотров

СОЧЕТАНИЯ ЖАНРОВ	РЕЙТИНГИ 1	2	3	4	5	Общий итог
Drama	620	1130	3238	4987	3282	13257
Comedy	929	1434	3023	2987	1455	9828
Comedy Romance	282	577	1560	1794	842	5055
Drama Romance	207	486	1233	1753	1088	4767
Action Thriller	201	435	1133	1228	553	3550
Drama Thriller	146	304	713	918	546	2627
Comedy Drama	148	249	682	877	466	2422
Drama War	47	119	414	745	687	2012
Action Adventure Sci-Fi	83	250	617	617	298	1865
Horror	193	262	448	430	225	1558
Action Adventure	106	167	383	494	382	1532
Animation Children's Musical	59	140	474	520	296	1489
Action Adventure Thriller	81	201	434	449	177	1342
Mystery Thriller	43	119	338	511	317	1328
Crime Drama	46	90	306	498	387	1327
Children's Comedy	199	239	409	238	69	1154
Thriller	89	159	337	341	168	1094
Action Adventure Romance Sci-Fi War	17	39	154	384	496	1090
Action Drama War	25	70	206	381	341	1023
Drama Sci-Fi	26	84	224	429	258	1021
Action Sci-Fi Thriller	59	99	223	374	255	1010
Action Romance Thriller	67	150	296	348	134	995
Drama Romance War	39	74	194	343	341	991
Crime Drama Thriller	34	69	210	326	264	903
Action	124	189	313	206	47	879
Crime Thriller	35	71	146	282	298	832
Horror Thriller	76	111	224	220	101	732
Comedy Sci-Fi	31	69	200	272	143	715

	Среднее по полю rating	Количество по полю rating
Animation Comedy Thriller	4,5	112
Animation	4,4	72
Sci-Fi War	4,3	194
Crime Film-Noir	4,3	4
Action Adventure Drama Romance Sci-Fi War	4,2	367
Action Adventure Romance Sci-Fi War	4,2	1090
Action Adventure Romance War	4,2	152
Adventure War	4,2	317
Action Drama Romance	4,2	428
Action Crime Drama	4,1	691
Film-Noir Mystery	4,1	211
Film-Noir Sci-Fi	4,1	275
Film-Noir	4,1	67
Film-Noir Romance Thriller	4,1	52
Horror Romance Thriller	4,1	239
Crime Film-Noir Mystery	4,1	40
Comedy Mystery Thriller	4,1	115
Adventure Children's Drama Musical	4,1	246
Film-Noir Mystery Thriller	4,0	161
Crime Film-Noir Mystery Thriller	4,0	354
Comedy Thriller	4,0	212
Drama Mystery	4,0	577
Drama Mystery Sci-Fi Thriller	4,0	259
Drama War	3,9	2012
Action Sci-Fi Thriller War	3,9	284
Action Adventure Comedy Romance	3,9	532
Comedy Mystery Romance Thriller	3,9	40
Action Drama War	3,9	1023