

Использование Microsoft Excel для анализа и обработки информации из Kaggle

Пастухов Никита Сергеевич

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

В статье рассматриваются возможности Microsoft Excel для анализа и обработки данных с платформы Kaggle. Описываются основные методы работы с данными, включая фильтрацию, сортировку, визуализацию и применение формул, что делает Excel удобным инструментом для первичного анализа.

Ключевые слова: Microsoft Excel, Kaggle, сводные таблицы, анализ данных, обработка данных, визуализация данных, формулы Excel.

Using Microsoft Excel to analyze and process information from Kaggle

Pastukhov Nikita Sergeevich

Sholom-Aleichem Priamursky State University

Student

Abstract

The article discusses the capabilities of Microsoft Excel for analyzing and processing data from the Kaggle platform. It describes the basic methods of working with data, including filtering, sorting, visualization, and applying formulas, which make Excel a convenient tool for primary analysis.

Keywords: Microsoft Excel, Kaggle, pivot tables, data analysis, data processing, data visualization, Excel formulas.

1 Введение

1.1 Актуальность

В современных условиях большие объемы данных требуют эффективных инструментов для их анализа, и Microsoft Excel остается одним из наиболее доступных и популярных решений. Платформа Kaggle предоставляет огромное количество открытых датасетов, но многие пользователи, особенно начинающие аналитики, сталкиваются с трудностями при их обработке. Данная статья актуальна, поскольку демонстрирует, как с помощью базовых и продвинутых функций Excel можно проводить анализ данных без сложного программирования

1.2 Обзор исследований

Статья В.А. Титова и С.Н. Цыганова посвящена применению математических методов и инструментальных средств, таких как автоматизированные системы управления стоимостью медицинских услуг и рабочее место экономиста, для повышения эффективности деятельности медицинских организаций, включая клиническую, организационную и экономическую составляющие [1]. Н.Н. Яроменко и А.С. Гореликова и Е.Н. Буркот исследуют тесную взаимосвязь между численностью занятого населения и среднегодовой заработной платой в Краснодарском крае с использованием регрессионно-корреляционного анализа, показывая, что увеличение заработной платы на 1% приводит к росту занятости на 0,105% [2]. О.М. Драпкина и Л.Ю. Волкова и Р.Н. Шепель и др. анализируют качество образовательных мероприятий, проведенных с использованием телемедицинских технологий, и выявляет, что содержание лекций, их практическая полезность и коммуникативные навыки лекторов являются ключевыми факторами удовлетворенности участников [3]. С.В. Исакова и Н.П. Орлянская описывают разработку автоматизированной системы в MS Excel для обработки биометрических данных кукурузы, включая расчёт площади листовой поверхности и фотосинтетического потенциала, что значительно упрощает анализ и прогнозирование результатов сельскохозяйственных исследований [4]. А.Г. Куделин исследует влияние информационных технологий на приверженность пациентов к терапии, показывая, что активные пользователи интернета (средний возраст 39,8 лет) демонстрируют более высокую комплаентность, но также повышенный уровень тревоги и депрессии [5]. Е.В. Конышев и Ю.А. Колесова и А.А. Веприкова исследуют особенности использования цифровых сервисов жителями Кировской области в туристических поездках, выявляя их консервативные предпочтения (например, доверие к официальным сайтам и низкий спрос на виртуальные экскурсии) и основные используемые платформы, такие как Avito, Яндекс.Карты и сайт РЖД [6]. А.В. Рахимбеков и Т.А. Адильханов и М.Р. Мадиева исследуют диагностические возможности диффузионно-взвешенной МРТ и измеряемого коэффициента диффузии (ADC) для дифференциации степени злокачественности менингиом головного мозга, показывая, что значения ADC значительно ниже у злокачественных форм (МIII) по сравнению с доброкачественными (MI и MII) [7].

1.3 Цель исследования

Целью данной статьи является демонстрация возможностей Microsoft Excel для анализа и обработки данных с платформы Kaggle, а также выявление эффективных методов работы с табличными данными.

Исследование направлено на то, чтобы показать, как стандартные инструменты Excel могут быть использованы для решения практических задач аналитики.

2 Материалы и методы

В исследовании использовался структурированный датасет с Kaggle, обработанный в Microsoft Excel с помощью инструментов импорта данных, формул, сводных таблиц, анализа данных и визуализации.

3 Результаты и обсуждения

Для начала, нужно скачать dataset с Kaggle и извлечь его. Возьмем, например: Netflix Movies and TV Shows (<https://www.kaggle.com/datasets/shivamb/netflix-shows>). Скачаем архив и распакуем его. Открываем csv файл, переходим во вкладку «Данные», нажимаем «Получить внешние данные из текста». Выбираем файл, в который хотим сохранить данные. Отмечаем кодировку UTF-8 и разделитель «Запятая». Сохраняем файл с расширением (*.xlsx).

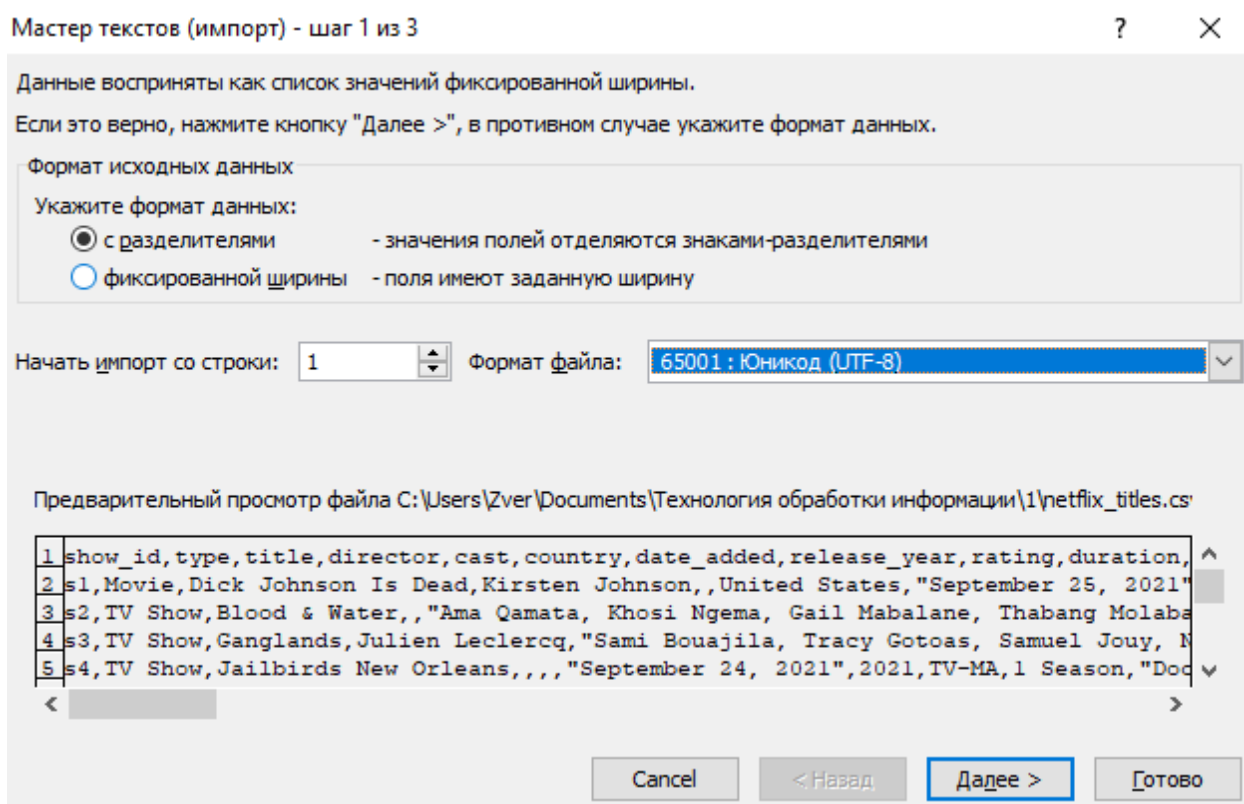


Рис 1 - Формат файла

Данный диалог позволяет установить разделители для текстовых данных. Результат выводится в окне образца разбора.

Символом-разделителем является:

- ☐ знак табуляции
☐ точка с запятой
☒ запятая
☐ пробел
☐ другой:

☐ Считать последовательные разделители одним

Ограничитель строк:

Образец разбора данных

show_id	type	title	director	cast
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	
s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail
s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas,
s4	TV Show	Jailbirds New Orleans		

Cancel

< Назад

Далее >

Готово

Рис 2 - Выбор разделителя

1. Анализ данных (сортировка, условное форматирование, сводная таблица, формулы)

1) Сортировка и фильтрация

Выделим шапку и нажмем «Сортировка и фильтр». Отсортируем по типу, стране и году выпуска (TV Show, Turkey, 2021).

A	B	C	D	E	F	G	H
show_id	type	title	director	cast	country	date_added	release_year
s704	TV Show	The Gift		Beren Saat, Mehmet Günsür, Metin Akdülger, Melisa Şenolsun, Başak Köklükaya, Civan Canova, Ti	Turkey	June 17, 2021	2021
s973	TV Show	Fatma		Burcu Biricik, Uğur Yücel, Mehmet Yılmaz Ak, Hazal Türesan, Gülçin Kütür Şahin, Olgun Tokar, Çağrı	Turkey	April 27, 2021	2021
s1374	TV Show	50M2		Engin Öztürk, Kürşat Alınçak, Cengiz Bozkurt, Aybüke Pusat, Tolga Tekin, Özgür Emre Yıldırım, Yiğit	Turkey	January 27, 2021	2021

Рис 3 - Сортировка и фильтрация

2) Условное форматирование

Закрасим только те значения столбца duration, где продолжительность составляет 2 сезона и 3 сезона.

Создание правила форматирования

Выберите тип правила:

- ▶ Форматировать все ячейки на основании их значений
- ▶ Форматировать только ячейки, которые содержат
- ▶ Форматировать только первые или последние значения
- ▶ Форматировать только значения, которые находятся выше или ниже среднего
- ▶ Форматировать только уникальные или повторяющиеся значения
- ▶ Использовать формулу для определения форматируемых ячеек

Измените описание правила:

Форматировать только ячейки, для которых выполняется следующее условие:

Значение ячейки между 2 Seasons и 3 Seasons

Образец: Формат не задан

Формат...

OK Отмена

Рис 4 - Условие форматирования

	A	B	C	D	E	F	G	H	I	J
1	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
2	11	Movie	Dick Johnson Is Dead	Kirsten Johnson	Ama Qamata, Khosi Ngema, Gail Mababane, Thabang Molaba, Dillon Windvogel, Natasha Thahane	United States	September 25, 2021	2020	PG-13	90 min
3	12	TV Show	Blood & Water		Julien Leclercq	South Africa	September 24, 2021	2021	TV-MA	2 Seasons
4	13	TV Show	Ganglands		Samir Bouajila, Tracy Gotoas, Samuel Jouy, Nabila Akkari, Sofia Lesaffre, Salim Kechiouche, Noureddine Farhi, Geert Van R	France	September 24, 2021	2021	TV-MA	1 Season
5	14	TV Show	Jailbirds New Orleans		Mayur More, Jitendra Kumar, Ranjan Raj, Alam Khan, Ahsaas Channa, Revathi Pillai, Urvil Singh, Anu	India	September 24, 2021	2021	TV-MA	1 Season
6	15	TV Show	Kota Factory		Mike Flanagan, Kate Siegel, Zach Gilford, Hamish Linklater, Henry Thomas, Kristin Lehman, Samantha Sloan, Igby Rigney, Rahul Kohli, An	United States	September 24, 2021	2021	TV-MA	2 Seasons
7	16	TV Show	Midnight Mass		Robert Cullen, Vanessa Hudgens, Kimiko Glenn, James Marsden, Sofia Carson, Lisa Koshy, Ken Jeong, Elizabeth Perkins, Jane Krakowski,	United States	September 24, 2021	2021	TV-MA	1 Season
8	17	Movie	My Little Pony: A New Generation		Haile Gerimo	Ghana	September 24, 2021	2021	PG	91 min
9	18	Movie	Sankofa			United States, Ghana	September 24, 2021	1993	TV-MA	125 min

Рис 5 - Результат условного форматирования

3) Сводная таблица

Щелкаем по таблице и выбираем во вкладке «Вставка», «Сводная таблица». Создадим сводную таблицу рейтинга фильмов и телешоу.

Сводная таблица

Название строки: **rating**

Название столбца: **rating**

Список полей сводной таблицы:

- ☐ show_id
- ☐ type
- ☐ title
- ☐ director
- ☐ cast
- ☐ country
- ☐ date_added
- ☐ release_year
- ☒ rating
- ☐ duration
- ☐ listed_in
- ☐ description

Прогнозируйте поля между указанными ниже:

Выбор отчетов: ☒ Вывести отчеты

Название строки: **title**

Название столбца: **rating**

Сводная таблица:

title	rating
11	PG-13
12	TV-MA
13	TV-MA
14	TV-MA
15	TV-MA
16	TV-MA
17	PG
18	TV-MA

Рис 6 - Сводная таблица рейтинга фильмов и телешоу

2. Формулы

1) Min/Max

Определим самый(-ое) старый(-ое) и самый(-ое) новый(-ое) фильм/телешоу.

Введем формулу =МИН() в ячейку O3 и диапазон Н:Н. Введем формулу =МАКС() в ячейку O4 и диапазон Н:Н.

1925
2021

Рис 7 - Результаты поиска

2) Комбинированные условия

Посчитаем сколько кино с 2020 годом выхода и с PG-13 рейтингом.

Для этого введем в ячейку O2 формулу =ЕСЛИ(И(B2="Movie";H2=2020;I2="PG-13");1;0), а в ячейку Q2 – формулу подсчета нужных нам строк: =СЧЁТЕСЛИ(O:O;1). Результат: 21

fx =СЧЁТЕСЛИ(O:O;1)						
	L	M	N	O	P	Q
	ife, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable.			1		21
	pe Town teen sets out to prove whether a private-school swimming star is her sister who was abducted at birth.			0		
	ul drug lord, skilled thief Mehdi and his expert team of robbers are pulled into a violent and deadly turf war.			0		
	own among the incarcerated women at the Orleans Justice Center in New Orleans on this gritty reality series.			0		
	to train India's finest collegiate minds, an earnest but unexceptional student and his friends navigate campus life.			0		

Рис 8 - Итог подсчета

3) Подсчет по условиям

Найдем кино/телешоу, которые были выпущены в 1945 году.

Впишем в ячейку O2 формулу: =СЧЁТЕСЛИ(H2:H8810;1945).

Результат: 4

fx =СЧЁТЕСЛИ(H2:H8810;1945)						
	L	M	N	O		
	filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable.					4
	town teen sets out to prove whether a private-school swimming star is her sister who was abducted at birth.					
	rug lord, skilled thief Mehdi and his expert team of robbers are pulled into a violent and deadly turf war.					
	wn among the incarcerated women at the Orleans Justice Center in New Orleans on this gritty reality series.					
	rain India's finest collegiate minds, an earnest but unexceptional student and his friends navigate campus life.					
	it brings glorious miracles, ominous mysteries and renewed religious fervor to a dying town desperate to believe.					

Рис 9 - Итог подсчета

3. Поиск слов, словосочетаний, предложений в тексте

Найдем кино, в описании которого содержатся слова: «...три брата, из всех сил стремящиеся стать честными, становятся мишенями преступного мира...».

Переходим во вкладку «Главная», затем «Редактирование», «Найти». И вставляем отрывок описания. Нажимаем «Найти все».

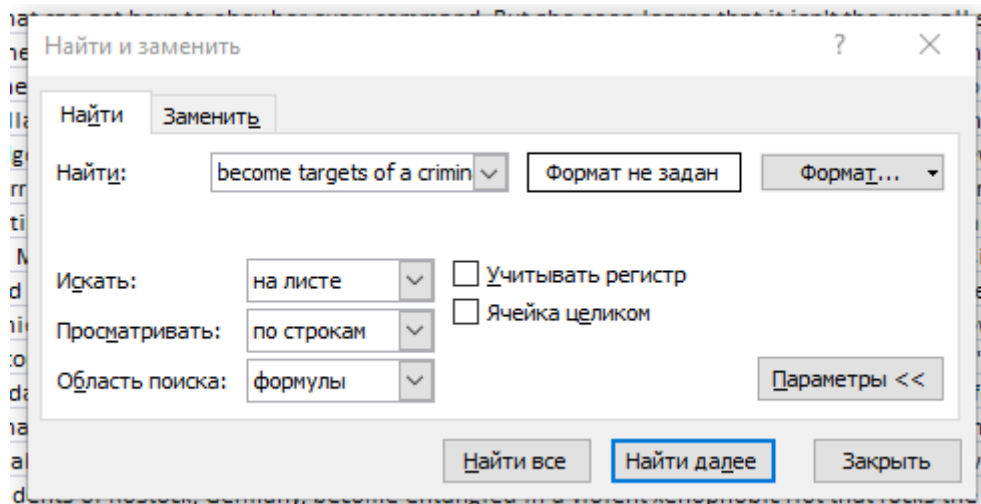


Рис 10 - Поиск отрывка в предложении

When three brothers struggling to go straight become targets of a criminal underworld, they use the making of a movie as a front to transport drugs.					
Н	И	Ж	К	Л	М
2017	TV-MA	54 min	Comedies, Music & Musicals, Romantic Movies	After Santa tells him he needs 75,000 new babies by Christmas to meet toy supply, Michael Bolton hosts a telethon to inspire the world to make love.	
2017	TV-MA	78 min	Stand-Up Comedy	Comedian Daniel Sosa inhabits various characters, including a young father returns home from jail eager to care for a young couple falls in love, but conflicting loyalties drive a revenge-seeking diamond thief gathers a ragtag crew	
2014	TV-MA	86 min	Dramas	A selfish, manipulative schoolteacher takes advantage of his special ops agency is wiped out, Spain hastily	
2015	TV-PG	154 min	Action & Adventure, Dramas, International Movies	A girl discovers a dog-training app that can get boys to	
2014	TV-14	179 min	Action & Adventure, Comedies, Dramas	Years after they were childhood sweethearts, Manuela	
2015	TV-14	93 min	Dramas, Independent Movies, International Movies	No-nonsense comic Bill Burr takes the stage in Nashville	
2016	TV-MA	97 min	Action & Adventure, Comedies, International Movies	A single mother, her boyfriend and her 8-year-old son try	
2016	TV-MA	108 min	Dramas, International Movies, Romantic Movies	When they flee Paris for a remote village in Greenland,	
2014	TV-Y	92 min	Children & Family Movies, Comedies	When shards of a cell phone get lodged in a teenager's	
2017	TV-MA	78 min	Stand-Up Comedy	After being raped in her home, a married woman chooses	
2015	TV-MA	87 min	Dramas, International Movies	The irrepressible Alonzo skewers Latino stereotypes, playing	
2016	TV-14	99 min	Comedies, International Movies	French comic Gad Elmaleh regales a Montreal crowd with	
2017	TV-MA	91 min	International Movies, Sci-Fi & Fantasy, Thrillers	A sexual assault victim, unjustly tried for the attempted	
2016	TV-MA	102 min	Dramas, International Movies	Two buddies working dead-end cashier jobs endure a	
2017	TV-MA	67 min	Stand-Up Comedy	Wicked one-liners and soul-baring confessions converge	
2017	TV-MA	57 min	Dramas, International Movies	DJ Icarus is at the top of his game, dazzling fans and	
2016	TV-14	130 min	Dramas, International Movies	Tasked with unraveling the assassination of Archduke	
2017	TV-MA	80 min	Comedies	Erik's peaceful rural family life is shaken by the arrival	
2017	TV-MA	66 min	Stand-Up Comedy	In 1992, three radically different residents of Rostock, Germany, become entangled in a violent xenophobic riot that rocks the beleaguered city.	
2008	TV-MA	105 min	Dramas, International Movies, Music & Musicals	The rise and smash success of Elite modeling agency founder John Casablancas is chronicled in this biography of the man who invented the supermodel.	
2014	TV-PG	98 min	Dramas, International Movies	A psychiatrist is plagued by flashbacks to a terrifying attack as she treats a new patient who was horribly disfigured in an accident.	
2014	TV-MA	91 min	International Movies, Thrillers	When three brothers struggling to go straight become targets of a criminal underworld, they use the making of a movie as a front to transport drugs.	
2014	TV-MA	123 min	Dramas, International Movies		
2016	TV-14	82 min	Documentaries, International Movies		
2017	TV-MA	104 min	Horror Movies, Thrillers		
2015	TV-MA	109 min	Comedies, Dramas, International Movies		

Рис 11 - Результат поиска

4. Визуализация данных

Представим в виде диаграммы весь телеконтент, предлагаемый Netflix.

Создадим табличку при помощи формулы =СЧЕТЕСЛИ(), и найдем сколько в перечне «Movie», «TV Show», и «William Wyler».

В ячейку P4 впишем: =СЧЁТЕСЛИ(B2:B8810;"Movie"). То же самое делаем с остальными. Делим найденные количества на 8809, переводим в проценты.

Далее, выделяем получившуюся табличку, переходим во «Вставка», «Диаграммы», «Круговая».

Для удобства подпишем проценты для каждой доли.

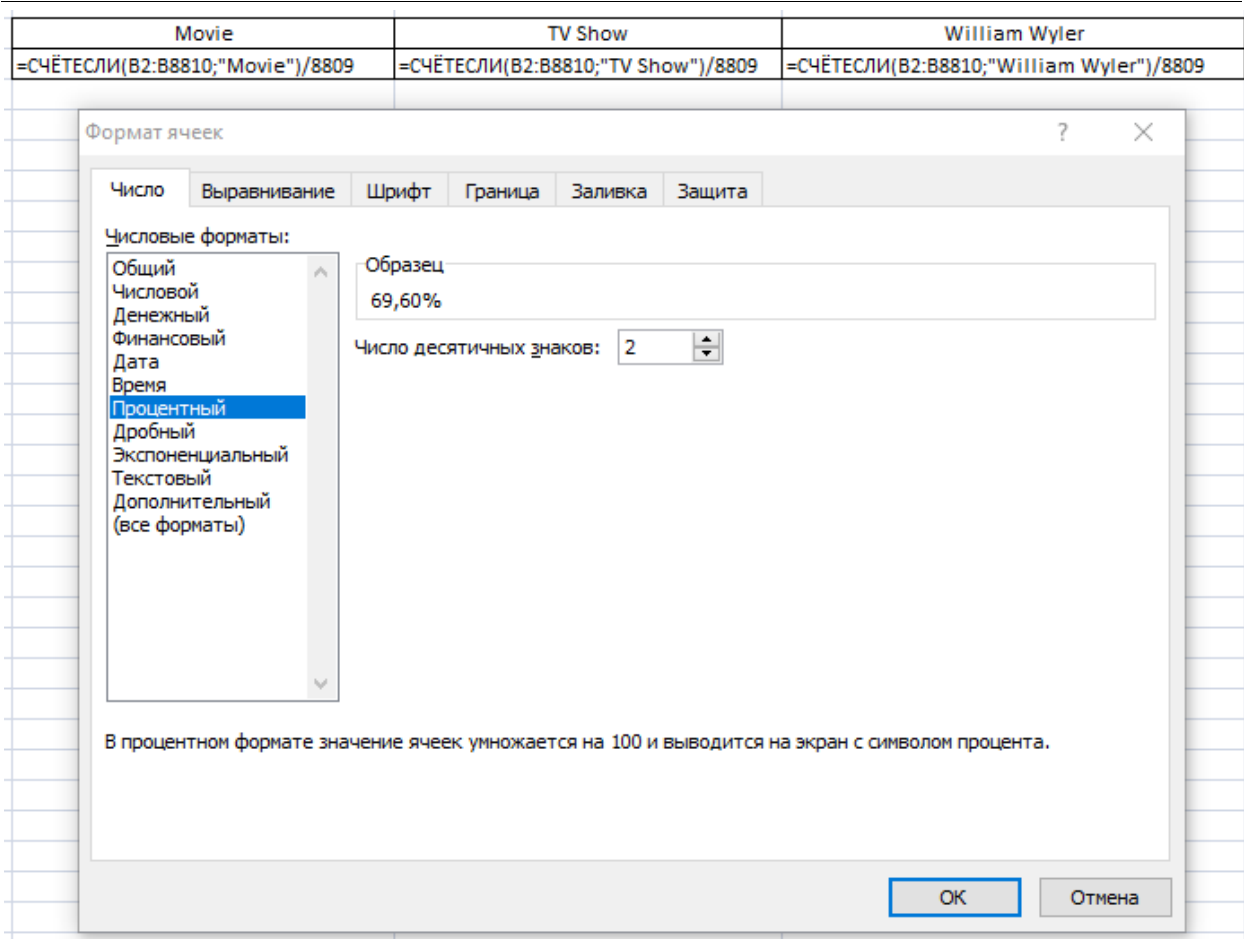


Рис 12 - Построение таблицы

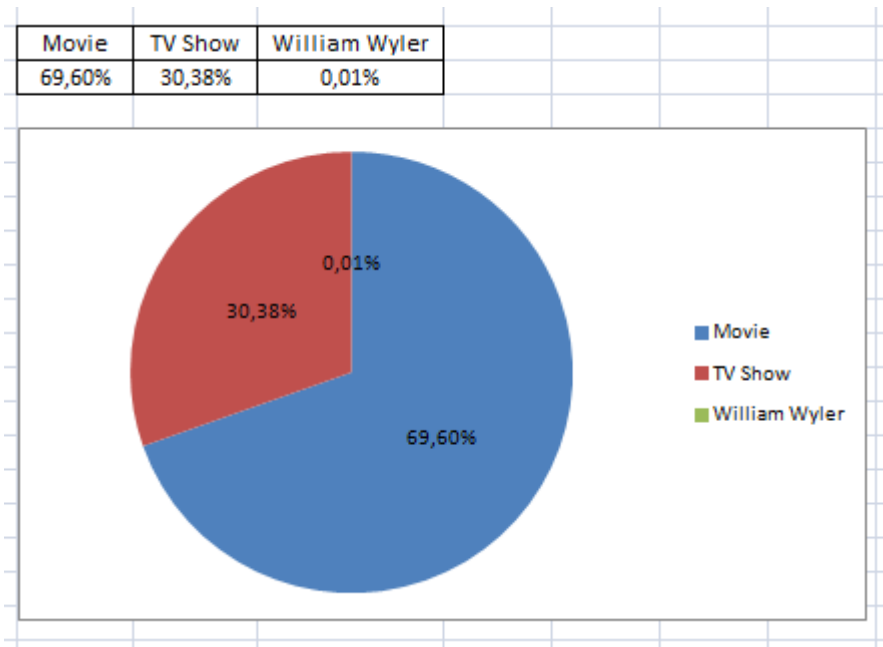


Рис 13 - Диаграмма

4 Выводы

Исследование подтвердило, что Microsoft Excel эффективен для анализа данных, предлагая удобные инструменты (сводные таблицы, формулы, визуализацию) для пользователей без навыков программирования. Хотя Excel

уступает специализированным платформам в обработке больших массивов, он остаётся ценным решением для быстрого анализа и образовательных целей, демонстрируя баланс между функциональностью и доступностью.

Библиографический список

1. Титов В.А., Цыганов С.Н. Математические методы и инструментальные средства повышения эффективности деятельности медицинских организаций // Открытое образование. 2016. Т. 20. № 6. С. 70-76. URL: <https://www.elibrary.ru/item.asp?id=27538904>
2. Яроменко Н.Н., Гореликова А.С., Буркот Е.Н. Эконометрический анализ зависимости между численностью занятого населения и среднегодовой заработной платой в краснодарском крае // Естественно-гуманитарные исследования. 2023. № 3 (47). С. 270-272. URL: <https://www.elibrary.ru/item.asp?id=54344893>
3. Драпкина О.М., Волкова Л.Ю., Шепель Р.Н., Жамалов Л.М., Астанина С.Ю., Ваховская Т.В. Анализ качества образовательных мероприятий, проведенных с использованием телемедицинских технологий // Кардиоваскулярная терапия и профилактика. 2022. Т. 21. № S3. С. 36-42. URL: <https://www.elibrary.ru/item.asp?id=49184902>
4. Исакова С.В., Орлянская Н.П. Применение компьютерных методов обработки данных при расчётах площади листовой поверхности растений кукурузы и ее фотосинтетического потенциала // Вестник Чувашской государственной сельскохозяйственной академии. 2019. № 4 (11). С. 52-60. URL: <https://www.elibrary.ru/item.asp?id=42773812>
5. Куделин А.Г. Анализ влияния информационных технологий на приверженность пациентов к терапии на амбулаторном этапе // Молодежный инновационный вестник. 2022. Т. 11. № S1. С. 129-131. URL: <https://www.elibrary.ru/item.asp?id=48970058>
6. Konyshchev E.V., Kolesova Yu.A., Veprikova A.A. The features of using digital services in tourist trips by citizens of kirov region (based on the results of 2022 survey) // Services in Russia and Abroad. 2023. Т. 17. № 2 (104). С. 141-152. URL: <https://www.elibrary.ru/item.asp?id=54117548>
7. Rakhimbekov A.V., Adylkhanov T.A., Madiyeva M.R., Kuanysheva A.G., Belikhina T.I., Raissov D.T., Baizakova M.N., Dyussyupova S.A. Diagnostic capabilities of diffuse-weighted magnetic resonance imaging in brain meningiomas // Science & Healthcare. 2019. Т. 21. № 3. С. 84-91. URL: <https://www.elibrary.ru/item.asp?id=38593816>