**Введение:**

**Добрый день! Меня зовут Мария и сегодня я расскажу вам о своем проекте по анализу базы данных лучших книг всех времен по мнению пользователей сайта GoodReads”.**

**· Good evening! My name is Maria and today I will tell you about my project on analyzing the database of the best books of all time according to users of the GoodReads website.**

**Слайд 1.**

**Набор данный содержит подробную информацию о каждой книге: начиная названием, автором и годом издания, заканчивая рейтингом, количеством оценок, ценой и количеством страниц.**

**This data set contains detailed information about each book: from the title, author and year of publication to the rating, price and amount of pages.**

**Цель и задачи:**

**Проанализировать тенденции в жанрах и ценах на книги.**

**Изучить демографический состав авторов.**

**Предоставить рекомендации, основанные на данных, для поддержки ключевых участников книжного рынка.**

**Purpose & Objectives:**

**To analyze trends in book genres and pricing.**

**Explore the demographic profile of authors.**

**Provide data-driven recommendations to support key stakeholders in the book industry.**

**Слайд 2: Проект и мотивация -- Slide 2: Project and Motivation**

* ***Analysis of the book market using data-driven insights.***
* ***Explore trends in book genres, author demographics, and pricing patterns***
* ***Project aim: Provide a comprehensive view of the literary world***
* ***Reveal hidden patterns and trends through data***
* ***Find valuable information for publishers, authors, and readers***

**The project focuses on analyzing the list of popular books, revealing trends in genres, authors demographics, and pricing patterns. The motivation for this project comes from the desire to understand desire to understand consumer preferences, pricing strategies, and gaps in the market.**

**Russian:**

**Проект посвящен анализу списка популярных, выявлению трендов в жанрах, демографии авторов и ценовых паттернов. Мотивация для этого проекта исходит из желания понять, предпочтения потребителей, ценовые стратегии, наличии пробелов на рынке, что поможет издателям и читателям принимать более информированные решения.**

**Слайд 3: Ограничения проекта // Slide 4: Project Limitations**

**One of the main challenges was working with a large and diverse dataset, which required significant computational resources. Additionally, dealing with incomplete or inconsistent data created difficulties in ensuring the accuracy of the analysis.**

**Одной из главных проблем было использование большого и разнообразного набора данных, что требовало значительных вычислительных ресурсов. Кроме того, работа с неполными или некорректными данными создавала трудности в обеспечении точности анализа.**

**Говоря подробнее о трудностях с которыми я столкнулась:**

1. **Временные ограничения: Нехватка времени для более глубокой обработки данных и использования внешних API.**
2. **Time Constraints: There was limited time to perform more extensive data analysis or to incorporate additional external data sources.**
3. **Технические ограничения: Размер базы данных и количество факторов, которые нужно учитывать, значительно увеличивали время обработки данных и требовали особого подхода к управлению ресурсами.**
4. **Technical Limitations: The size of the dataset and the number of factors involved required careful resource management and optimization strategies during the data processing phase.**

**I overcame computational difficulties by optimizing data processing methods and focusing on key parameters.**

**Я преодолела вычислительные сложности, оптимизировав методы обработки данных и сосредоточив внимание на ключевых параметрах.**

**Слайд 4: Инструменты и стек технологий ///Slide 4: Tools and Technology Stack**

* + **В рамках этого проекта я использовала несколько инструментов и технологий для эффективного управления и выполнения задач. Для управления проектом и отслеживания задач я использовала Trello, что помогло мне организовать процесс работы и отслеживать прогресс на всех этапах.**
  + **As part of this project, I utilized several tools and technologies to ensure smooth management and execution. For project management and task tracking, I used Trello, which helped me keep everything organized and track progress throughout the project.**
  + **Для разработки я работала в PyCharm, где проводила фильтрацию данных, удаление дубликатов, форматирование дат и обработку отсутствующих значений.**
  + **For development, I worked in PyCharm, where I performed data filtering, removed duplicates, formatted dates, and handled missing values.**
  + **Основной анализ и предобработку данных я выполнила в Python с использованием библиотек pandas и NumPy. Эти инструменты позволили мне определить пол авторов, извлечь основные жанры, провести анализ распределения оценок и изучить тренды популярности жанров, цен книг и количества страниц.**
  + **The main data analysis and preprocessing were done in Python with the help of libraries like pandas and NumPy. These tools allowed me to detect the gender of authors, extract main genres, analyze the distribution of ratings, and examine trends in genre popularity, book prices, and page counts.**
  + **Я также интегрировала Open Library API, чтобы дополнить и верифицировать данные, заполняя пропуски или недостающую информацию.**
  + **I also integrated the Open Library API to validate and enrich the dataset by filling in missing or incomplete information.**
  + **Для контроля версий и удобного взаимодействия я использовала Git и GitHub, что обеспечило отслеживание всех изменений и эффективность разработки.**
  + **For version control and collaboration, I used Git and GitHub, ensuring that all changes were tracked and development was efficient.**
  + **И наконец, я создала интерактивный дашборд в Tableau, который позволяет пользователям исследовать данные с помощью различных фильтров и получать полезные инсайты с разных точек зрения."**
  + **Finally, I used Tableau to create an interactive dashboard, which allows users to explore the data through various filters and gain insights from different perspectives."**

**Слайд 5: Корреляционная матрица (Python Visualization)//Slide 5: Correlation Matrix**

**"На этом слайде я хочу обратить внимание на визуализации, которые были созданы в Python, и особенно на представленную здесь корреляционную матрицу. Эта матрица помогает нам понять взаимосвязи между различными числовыми характеристиками в наборе данных, такими как рейтинги, количество страниц, цена и метрики популярности, такие как bbeScore и bbeVotes.**

**In this slide, I want to highlight the visualizations created using Python, particularly focusing on the correlation matrix displayed here. As you can see, this matrix helps us understand the relationships between different numerical features in the dataset, such as ratings, number of pages, prices, and popularity metrics like bbeScore and bbeVotes.**

**Одним из ключевых выводов является сильная положительная корреляция между рейтингом и процентом положительных оценок (LikedPercent). Это указывает на то, что книги с более высокими рейтингами, как правило, имеют более высокий процент читателей, которым они понравились. Мы также видим умеренную корреляцию между количеством оценок (numRatings) и метриками популярности (bbeScore и bbeVotes), что говорит о том, что книги с большим количеством оценок, как правило, более популярны.**

**One of the key takeaways from this matrix is the strong positive correlation between Rating and LikedPercent. This indicates that books with higher ratings tend to have a higher percentage of readers who liked them. We also observe a moderate correlation between the number of ratings (numRatings) and popularity scores (bbeScore and bbeVotes), suggesting that books with more ratings tend to be more popular.**

**Интересно отметить, что нет значительной корреляции между количеством страниц или ценой книг и такими показателями, как рейтинги или популярность. Это означает, что длина книги и её цена не обязательно влияют на её успех в терминах рейтингов или популярности.**

**Interestingly, there seems to be no significant correlation between the number of pages or the price of the books with other metrics like ratings or popularity. This tells us that book length and price don't necessarily influence a book's success in terms of ratings or its overall popularity.**

**Эту матрицу я создала в Python в процессе исследования данных, и она предоставила ценные сведения о том, как различные факторы взаимодействуют друг с другом.**

**I created this matrix in Python as part of the data exploration process, and it provided valuable insights into how different factors interact with each other."**

****

**Слайд 8: Дашборд Tableau//Slide 8: Dashboard Demo in Tableau**

**Дашборд 1:**

* **Временная шкала публикации книг по полу автора:**
  + **Этот график иллюстрирует временную шкалу публикаций книг, разделенных по полу автора. Видно, что как мужчины, так и женщины-авторы увеличили количество публикаций, особенно с середины XX века.**

**Book Publication Timeline by Gender:**

* **This graph illustrates the timeline of book publications divided by author gender. It shows that both male and female authors have seen growth in publications, with notable increases from the mid-20th century onward.**
* **Среднее количество страниц по жанрам:**
  + **Такие жанры, как Биография и Историческая литература, в среднем имеют большее количество страниц по сравнению с другими жанрами. Это может свидетельствовать о том, что книги в этих жанрах часто содержат более глубокое содержание или развернутые повествования.**

**Average Pages by Genre:**

* **Genres like Biography and Historical Fiction generally have a higher average page count compared to other genres. This suggests that books in these genres often provide more in-depth content or narratives.**
* **Цена vs. Количество страниц с линиями тренда по жанрам:**
  + **Точечный график с линиями тренда показывает, что цена книги не всегда коррелирует с количеством страниц. В некоторых жанрах, таких как Романтика и Художественная литература, цена остается стабильной независимо от длины книги, в то время как в других жанрах наблюдается большая вариативность.**
* **Price vs. Pages with Trend Lines by Genre:**
  + **The scatter plot with trend lines shows that the price of a book does not always correlate with the number of pages. Some genres, like Romance and Fiction, have consistent pricing regardless of page length, while others show more variability.**

**Линии тренда на графике показывают направление и общую тенденцию данных, а также их зависимость. В контексте диаграммы, они дают представление о том, как одна переменная изменяется в зависимости от другой.**

**Например:**

**Trendlines on a chart show the direction and general trend of the data, as well as their relationship. In the context of a chart, they provide insight into how one variable changes in relation to another.**

**Если линия тренда направлена вверх, это означает положительную корреляцию — с увеличением одной переменной увеличивается и другая.**

**If the trendline is pointing up, it indicates a positive correlation - as one variable increases, the other increases.**

**Если линия тренда направлена вниз, это указывает на отрицательную корреляцию — по мере увеличения одной переменной, другая уменьшается.**

**If the trendline is pointing down, it indicates a negative correlation - as one variable increases, the other decreases.**

**Плоская линия тренда показывает отсутствие значимой зависимости между переменными.**

**В данном случае линии тренда помогают визуально понять, как, например, цена книги изменяется в зависимости от количества страниц по разным жанрам. Если линии жанров сильно различаются, это может указывать на то, что для некоторых жанров цена больше зависит от длины книги, чем для других.**

**For example:**

**A flat trendline shows that there is no significant relationship between the variables.**

**In this case, trendlines help us visualize how, for example, the price of a book changes depending on the number of pages in different genres. If the genre lines vary greatly, it may indicate that for some genres, the price is more dependent on the length of the book than for others.**

* **Количество книг по полу автора:**
  + **Гистограмма показывает, что мужчины-авторы публикуют больше книг по сравнению с женщинами-авторами, что может отражать историческую тенденцию в индустрии или предвзятость в сборе данных.**
* **Number of Books by Author Gender:**
  + **The bar chart reveals that male authors have published more books compared to female authors, indicating a possible historical trend in the industry or data collection bias.**

**Дашборд 2:**

* **Рейтинг vs. Процент понравившихся:**
  + **Существует сильная корреляция между рейтингом книги и процентом людей, которым она понравилась. Это говорит о том, что книги с более высокими рейтингами, как правило, более популярны и лучше воспринимаются читателями.**

**Rating vs. Liked Percent:**

**There is a strong correlation between the rating of a book and the percentage of people who liked it. This suggests that higher-rated books are generally more popular and well-received by readers.**

**Анализ графика "Count of Ratings by Genre"**

* **Classics:**
* **Жанр «Classics» имеет наибольшее количество оценок, что может свидетельствовать о его значительной популярности и влиянии на читателей. Классические произведения часто входят в школьные программы и рекомендуемые списки литературы, что также повышает количество их оценок.**

**Classics:**

**The "Classics" genre has the highest number of ratings, which may indicate its significant popularity and influence on readers. Classics are often included in school curricula and recommended reading lists, which also increases their ratings.**

* **Nonfiction:**
  + **«Nonfiction» может свидетельствовать о популярности образовательных книг.**
* **Fiction:**
  + **«Fiction» занимает второе место по количеству оценок. Это указывает на высокий интерес к художественной литературе, которая охватывает широкий спектр тем и стилей.**

**Nonfiction:**

**"Nonfiction" may indicate the popularity of educational books.**

**Fiction:**

**"Fiction" has the second highest number of ratings. This indicates a high interest in fiction, which covers a wide range of topics and styles.**

* **Fantasy:**
  + **Жанры «Fantasy» также показывают значительные значения. «Fantasy» часто имеет преданную фанатскую базу.**
* **Other Genres:**

**Остальные жанры, такие как «Romance», «Poetry», Grafic novels имеют меньшее количество оценок, что не обязательно говорит о их низкой популярности. Это может быть связано с тем, что аудитория этих жанров более нишевая или специфическая.**

**Fantasy:**

**"Fantasy" genres also show significant values. "Fantasy" often has a dedicated fan base.**

**Other Genres:**

**Other genres such as "Romance", "Poetry", and "Graphic novels" have fewer ratings, which does not necessarily indicate their low popularity. This may be due to the fact that the audience for these genres is more niche or specific.**

* **Интерпретация:**

**Этот график показывает общее количество оценок по каждому жанру, что можно интерпретировать как индикатор популярности жанра. Однако стоит учитывать, что количество оценок не всегда равно реальной популярности или качеству книг. Например, новинки могут иметь меньше оценок, несмотря на высокий интерес.**

**Interpretation:**

**This graph shows the total number of ratings for each genre, which can be interpreted as an indicator of the genre's popularity. However, it is worth keeping in mind that the number of ratings does not always equal the actual popularity or quality of the books. For example, new releases may have fewer ratings despite being highly popular.**

* **Средний BBE Score по жанрам:**
  + **Показатель BBE Score измеряет популярность или значимость книги.**

**Average BBE Score by Genre:**

* **The BBE Score measures a book's popularity or significance. Classics stand out with the highest average BBE Score, followed by Fantasy and Fiction reflecting their cultural impact and reader engagement.**

**Analysis of the "Count of Ratings by Genre" graph**

**Слайд 9: Next steps**

1. **Project Steps:**
   1. **Deeper Data Analysis: Conduct more advanced data analysis to uncover hidden patterns and insights within the dataset.**
   2. **API Integration: Integrate additional APIs to enrich the dataset, ensuring more accurate and comprehensive results.**
2. **Career Development Steps:**
   1. **Learning New Skills: Focus on learning new tools and technologies relevant to data analysis and visualization.**
   2. **Improving Existing Skills: Enhance proficiency in current technologies like Python, Tableau, and data processing libraries to handle more complex projects.**
   3. **Networking and Job Search: Actively engage in networking and apply for data analysis roles, utilizing skills gained from this project.**

**Columns**

**File: books\_1.Best\_Books\_Ever.csv**

**Title The title of the book. (String)**

**Series The series the book belongs to, if any. (String)**

**Author The author of the book. (String)**

**Rating The average rating of the book on GoodReads. (Float)**

**isbn The ISBN number of the book. (String)**

**genres The genres the book falls into. (String)**

**pages The number of pages in the book. (Integer)**

**publisher The publisher of the book. (String)**

**publishDate The date the book was published. (Date)**

**firstPublishDate The date the book was first published. (Date)**

**awards Any awards the book has received. (String)**

**numRatings The number of ratings the book has received. (Integer) //how many people rated the book**

**numRatings Количество оценок, полученных книгой. (Целое число)**

**ratingsByStars The ratings the book has received broken down by star ratings. (String)**

**ratingsByStars Рейтинги, полученные книгой, разбитые по звездам. (Строка)**

**likedPercent The percentage of readers who liked the book. (Float)**

**likedPercent Процент читателей, которым понравилась книга. (Число с плавающей точкой)**

**bbeScore The score of the book on GoodReads' Best Books Ever list. (Float)**

**bbeScore Оценка книги в списке лучших книг GoodReads. (Число с плавающей точкой)**

**bbeVotes The number of votes cast for the book on GoodReads' Best Books Ever list. (Integer)**

**bbeVotes Количество голосов, отданных за книгу в списке лучших книг GoodReads. (Целое число)**

**price The price of the book. (Float)**