End-to-End Mini Project (Paula Boks)

1. Dataset: the dataset can be found here on Kaggle, it was downloaded on January 9th 2025.

2. Questions and answers about the dataset (using sql queries, script can be found in this repo)

   i. **General**:
      a. How many pages do the books have on average?
      b. Which languages are the books in?
      c. Which genres are there?
      d. Which type of binding do the books have?
      e. What is the average rating of the books?
      f. What is the highest price of a book?

   ii. **Authors**:
      a. Which authors have more than one book?
      b. Which authors wrote the book with the highest rating?

   iii. **Highly rated books**:
      a. How many pages does the book with the highest rating have?
      b. Which genres do the book with a rating higher than 4 have?
      c. What are the five leading books in the "Best Books Ever" list?

3. Report on dataset (using Python)

   a) Structure and General Content

   - The dataset contains **25 columns** and over **52,000 rows**.
   - The rows each describe a specific **book**, and the columns are **information on this book** from the website goodreads.com.
   - It has a unique identifier (**primary key**) with the column "bookId".
   - **Other columns** contain the following information, among others: title, author, rating, description, language, genre, isbn (which is a second unique identifier but there are missing values), number of pages, publishing date, format of the book.
   - There are **several columns regarding rating**: the main rating, the number of ratings, the rating by stars and more rating concerning the "Best Books Ever" list.

   b) State of the data

   - **Missing values**: There is many NaNs especially in four columns: edition (over 90%), series (more than 55%), firstPublishDate (over 40%) and price (over 27%). The other columns have either none missing values or less than 1%.
   - There seem to be **more missing values in "isbn"** than found out using isna().sum() because "9999999999999" seems to stand for NaN.

- The columns containing information on **dates** are in different formats: parts are like this (04/28/09) and many others like this ('October 17th 2006').
- Many columns contain text ins **list format** (Python list format).

4. Cleaning and Feature Engineering

a) Cleaning steps

- For the EDA and SQL project, only the **following columns** are kept: "bookId", "title", "author", "rating", "language", "genres", "bookFormat", "pages", "awards", "bbeScore", "price".
- All rows with **NaNs** in it were dropped. Also "hidden NaNs" in "genres" were dropped (encoded as empty lists).
- All **duplicates** were dropped.
- **Datatypes** were corrected.
- **Cleaning individual columns**:
    - "**author**": For visualisation reasons, I deleted the additional info on whether it is a "Goodreads Author".

b) Feature engineering

- Column "genres": was encoded as lists and many had more than one allocation. I created a **new column "genre"** with only the firstly mentioned genre.
- Column "awards": like "genres", encoded in lists. I created a **new column "number_awards"** and counted the entries in the original column.
- Column "author": sometimes contained more than one author, or the name of the translater i.e. I created a **new column "main_author"** which solely contains the first mentioned author.

5. EDA

- In the dataset, there now are: **four numerical columns** (pages, bbeScore, price and number_awards) and **seven categorial columns** (bookId, title, author, rating, language, bookFormat and genre).

a) Distribution and correlation of numerical columns

- "rating" shows a **negative skew** and a **unimodal distribution** narrowly centered around a **value near 4**.
- "pages", "bbeScore", "price" and "number_awards" show a **highly positively skewed** distribution with **most values at the very low** end but **outliers** at higher values.
- There is only **very low correlation** between the columns, the strongest being 0.20 between number_awards and bbeScore (linear) and also 0.20 between rating and price (monotonic).

b) Frequency Counts of Categorial columns

- **Nora Roberts** has the most books in the list, followed by **James Patterson**, **Agatha Christie** and **Stephen King**.

- **English** is by far the leading language, followed by **French**, **Spanish** and **German**.
- The most prominent format is the **paperback**, followed by a hardcover. Kindle is already on 4th place.
- The most current genre is **fiction**, followed by **fantasy** and **young adult**.

c) Categorials vs. Numericals

- Some books have **very high prices** (147 books over €150, the highest almost €900)
- **Nora Roberts** has the **most ratings**, followed by **James Patterson**, **Agatha Christie** and **Stephen King**; yet **other authors** have the **highest average ratings**: 9 of them have a mean rating of 5 (highest score).
- **Alchemy** is the genre with the highest mean rating (4.65 out of 5), followed by **Baha I** (a religion) and **Dinosaurs**. Out of the ten most common genres **History** receives the highest mean rating (4.09).
- **19th century** is the genre with the highest mean price (€285.93), followed by **Comic Books** (€229.64) and **Apocalyptic** (€173.48). Out of the ten most common genres **History** has the highest mean price (€11.77).
- **Stephen King** received the most awards (97), followed by Neil Gaiman (75), China Mieville (69) and Suzanne Collins (62). Yet the book title with the most awards (41) is "**Hunger Games**" by Suzanne Collins, followed by "**Escape from Mr. Lemoncello's Library**" by Chris Grabenstein (27) and "**Twilight**" by Stephenie Meyer (26).

6. Inferential Statistics

a) Chi^2 and Cramer's V tests on categorial variables

- The only strong relationships I detected are (not surprisingly) between "author" and "genre" and (only slightly weaker) between "author" and "rating".