Data Overview

- **Citation/Contact Info:** Linkedin Profile
- **Created by** Atharva Inamdar on 9/15/2022
- **OG purpose of the data:** To predict the genre of a book based on its synopsis

Motivation

- **Who created it?:** Atharva Inamdar, a Data Scientist at DataEaze Systems Pvt. Ltd.
- **Why was it created?:** Inamdar noticed that sites like Goodreads and Amazon often rely on their users to categorize books into specific genres. He was interested in streamlining this process through automation.
- **Stakeholders:** Readers, data scientists, students, developers

Composition

- **Observations/Instances, i.e., rows documented by the set Biases of the metrics**
  There are instances where some records have the same titles but different summaries/genres. However, all entries are holistically unique.
- **Summary statistical overview of the data types**
  The data types include objects represented as strings for title, summary, and genre. There are 4657 columns of book records, with 4296 unique titles, 4542 unique summaries, and 10 unique genres. The genres include thriller, fantasy, science, history, horror, romance, crime, psychology, sports, and travel

Collection Process

- **How was it collected?** Inamdar collected data from the Goodreads website using Selenium. The top books from each respective genre page were collected.
- **Sampling method, if applicable:** Data was sampled from Goodreads using Selenium
- **Timeframe:** Updated a year ago, said to be updated regularly on a weekly basis.

Processing / Labeling

- **What processing work, if any?** Goodreads data was processed through Selenium, a python extension used to extract data from websites.
- **OG data available, if this is a processed version?** -unavailable (although possibly found uncompiled on goodreads)

Uses

- **What were its original intended Uses?** Expedite the organization of books by genre using machine learning. Inamdar aso encourages others to use his data to create a model which can predict the rating of the book based on other attributes in the collected data.
- **How should it not be used?** - not specified

Distribution

- **How will it be distributed?** Distributed to Kaggle users who have free access to csv
- **Licensing / Intellectual Property:** The project is public domain
- **Regulation:** Free use

Maintenance

- **Who will maintain it?** Atharva Inamdar
- **Handling errata?** Atharva Inamdar

- **Contribution process?** The project's top contributors are Swetanshu Goel, Prathamesh Gadekar and Deblina Ghosh.