Data Collection and Preprocessing Phase

Date	25 Jan 2025
Team ID	740678
Project Title	Amazon Kindle Store Reviews analysis
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	The goal of this project is to analyze Amazon Kindle store customer reviews to extract sentiment, identify common themes, and assess user satisfaction across different book categories. Using natural language processing (NLP) techniques, the analysis will help authors and publishers understand customer feedback and improve offerings.
Data Collection Plan	The data collection plan involves scraping or downloading publicly available customer reviews from the Amazon Kindle store. This includes review text, star ratings, reviewer metadata, review dates, and associated book information. The dataset will support sentiment analysis, topic modeling, and rating correlation.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kindle Book Reviews 1	Reviews from top-selling Kindle fiction books	Amazon Kindle Store – Fiction	CSV	200MB	Public Web Scraped Data

Kindle Book Reviews 2	Reviews from non-fiction Kindle books in self-help and business categories	Amazon Kindle Store – Non-fiction	CSV	180MB	Public Web Scraped Data
Metadata Set	Contains book metadata like genre, title, author, publication date	Aggregated from Amazon product pages	JSON	50MB	Public Web Scraped Data
Review Snapshot Sample	A manually collected sample of 100 reviews with full text and annotations for model testing	Internal cloud drive	Excel	1MB	Private (internal access only)