

Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1720184497
Project Title	Cereal Analysis Based on Ratings by using Machine Learning Techniques
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 project aims to predict cereal ratings based on nutritional information and reviews using machine learning techniques.
Data Collection Plan	Data will be collected from public datasets available on Kaggle, consumer review platforms, and the USDA nutritional database.
Raw Data Sources Identified	The raw data sources include the cereal dataset from Kaggle, e-commerce reviews, and USDA nutritional data.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Cereal Dataset	A comprehensive dataset of cereals including nutritional information and ratings.	https://www.kaggle.com/crawford/80-cereals	CSV	18 KB	Public
Amazon Reviews	Consumer reviews and ratings of various cereal brands from Amazon.	https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products	CSV	350 MB	Public (with Kaggle account)
USDA Nutritional Data	Detailed nutritional information from the USDA database.	https://fdc.nal.usda.gov/download-datasets.html	CSV	Varies	Public