CDS513: Predictive Business Analytics
Academic Session: Semester 2, 2022/2023
School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Appendix A – Topic Selection

No.	Name of Dataset	Description	Sample Size	Recommender
1.	H&M Personalized Fashion Recommendations	H&M Group is a family of brands and businesses with 53 online markets and approximately 4,850 stores. Our online store offers shoppers an extensive selection of products to browse through. But with too many choices, customers might not quickly find what interests them or what they are looking for, and ultimately, they might not make a purchase. To enhance the shopping experience, product recommendations are key. More importantly, helping customers make the right choices also has positive implications for sustainability, as it reduces returns, and thereby minimizes emissions from transportation.	>100000	X
		https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data		
2.	The Movies Dataset	These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been	45,000	X

ASSIGNMENT 1

		obtained from the official GroupLens website.		
		https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=movies_metadata.csv		
3.	Book-Crossing Dataset	The Book-Crossing (BX) dataset was collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. It contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.	278,858	X
		http://www2.informatik.uni-freiburg.de/~cziegler/BX/		
4.	Walmart Recruiting: Trip Type Classification	This dataset is used to classify shopping trips. Potential for market basket and recommender.	4521127	X
		https://www.kaggle.com/c/walmart-recruiting-trip-type-classification/data		
5.	Retail rocket recommender system dataset	The purpose of publishing is to motivate researchers in the field of recommender systems with implicit feedback.	>200000	X
		https://www.kaggle.com/retailrocket/ecommerce-dataset		
6.	Restaurant Data with Consumer Ratings	This dataset was used for a study where the task was to generate a top-n list of restaurants according to the consumer preferences and find the significant features.	>1100	X
		https://www.kaggle.com/datasets/uciml/restaurant-data-with-consumer-ratings?select=usercuisine.csv		
7.	Goodbooks-10k	This dataset contains ratings for ten thousand popular books. As to the source, let's say that these ratings were found on the internet. Generally, there are 100 reviews for	10000	X

ASSIGNMENT 1

		each book, although some have less - fewer - ratings. Ratings go from one to five.		
		https://www.kaggle.com/datasets/zygmunt/goodbooks-10k?select=ratings.csv		
8.	IMDB data from 2006 to 2016	Here's a data set of 1,000 most popular movies on IMDB in the last 10 years. The data points included are: Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascrore	1000	X
		https://www.kaggle.com/PromptCloudHQ/imdb-data		
9.	Anime Recommendations Database	This data set contains information on user preference data from 73,516 users on 12,294 anime. Each user is able to add anime to their completed list and give it a rating and this data set is a compilation of those ratings.	12294	X
		https://www.kaggle.com/CooperUnion/anime-recommendations-database		