CS 418: Introduction to Data Science
Final Project Proposal
Fall 2019

Project Team : Anusha Voloju, Maleeha Ahmed, Priyan Sureshkumar

Problem Specification :

The goal of this project is to analyse the important factors that make a book more popular and most read compared to others and to use this analysis to predict the average rating of a book and to categorize the books into corresponding genres.

Data Science Solution :

Data preprocessing is performed on the raw data to clean the observations and using Descriptive Statistics the importance of each variable is determined.
These conclusions on the variables are used to build a regression model for the prediction of average rating of the books.
Using classification or clustering techniques, the books are categorized into specific genres.

Description of Data Sources :

We are using Publishers dataset in CORGIS datasets. The dataset consists of information about different titles spanned across different genres and publishers. It also includes the average rating and overall sales ranking of amazon kindle store.

The variables in the dataset include:
Genre
Sold_by
Daily_average_amazon_revenue
Daily_average_author_revenue
Daily_average_gross_sales
Daily_average_publisher_revenue
Daily_average_units_sold
Publisher_name
Publisher_type
Average_rating
Sale_price
Sales_rank
Total_reviews

Source :

https://think.cs.vt.edu/corgis/csv/publishers/