**Clustering on Book Data**

**Group Name: Group Rohan Aryan**

**Group Members:**

|  |  |  |
| --- | --- | --- |
| **First name** | **Last Name** | **Student number** |
| **Aishlee** |  | **C0913045** |
| **Jayachandhran** | **Saravanan** | **C0910392** |
| **Prashanta** | **Timsina** | **C0916254** |
| **Rohan** | **Aryan** | **C0912902** |
| **Yogita** | **Sharma** | **C0913011** |

**Submission date: *17-MAR-2024***

Contents

[**Abstract** 2](#_Toc161566754)

[**Introduction** 3](#_Toc161566755)

[**Methods** 3](#_Toc161566756)

[Fetching Data 3](#_Toc161566757)

[Data Wrangling and EDA 4](#_Toc161566758)

[Natural Language Processing 8](#_Toc161566759)

[Encoding 10](#_Toc161566760)

[Outliers 11](#_Toc161566761)

[Learning: 12](#_Toc161566762)

[Handling outliers detection 13](#_Toc161566763)

[Clustering 14](#_Toc161566764)

[TF-IDF COSINE 17](#_Toc161566765)

[**References** 19](#_Toc161566766)

[Dataset and Onedrive links 20](#_Toc161566767)

# **Abstract**

**This report enclosed the detailed analysis of data scrapped from the e-commerce book website and the data is processed to create valid clusters based on the difference and similarity. It also compares various approaches to handle the outliers present in the data and the clear observation is posted with the description of the impact it created. Later, K-means and DBSCAN are applied to various combinations of data. The best results of the K-means and DBSCAN are pointed. The cosine similarity-based description clustering is carried and termed to be the best method for this activity, the plots are not clear so t-SNE dimension reduction is applied for clear interpretation. The in-depth analysis and challenges are discussed further along the methods of the document. The API method to scrap the data is also discussed in detail.**

**Keywords: clustering, K-means, DBSCAN, scrapping, API, NLP, Cosine similarity, TF-IDF and t-SNE**

# **Introduction**

In this project, we tried to work on an unsupervised machine learning model technique applied to the custom scraped data from a book website. Throughout this activity, we tried to touch on various data handling and processing methods that are common in the industry/research to apply to the collected raw data. Further, we applied basic Natural Language Processing (NLP) steps to the collected text data to emphasize the implementation of text processing by our team. Finally, we implemented two main clustering types on the processed data to categorize the unlabeled data into particular classes based on similarity and difference which we will discuss in this report.

The following are the team split and the role assigned to each one for the members:

|  |  |
| --- | --- |
| Data Scraping | Jayachandhran Saravanan  Rohan Aryan |
| Basic EDA and Data Handling | Aishlee  Yogita Sharma |
| Data Engineering and validation | Rohan Aryan  Jayachandhran Saravanan |
| NLP and clustering techniques | Prashanta Timsina  Rohan Aryan and  Jayachandhran Saravanan |
| Result validation and fine tuning | Jayachandhran Saravanan  Prashanta Timsina  Rohan Aryan |
| Documentation and Inference | Rohan Aryan and Jayachandhran Saravanan |

# **Methods**

## Fetching Data

We approached this activity by brainstorming various ideas and possible resource websites to scrap the data. Out of which, we choose to go with “eBooks” website to scrap our data on the particular subcategory called “computer A screenshot of a computer

Description automatically generatedscience” present along 34 similar categories. We find the dataset interesting as it will be challenging to get various fields of data which may contain text, integer, and Date Time formats Also, we wanted to understand the book or article recommendation system available in Kindle and medium blog post.

*Figure1: website*

We explored three scrapping tools that are easy to configure and work with HTML content. Application Programming Interface-based web scraping is utilized in this project, where the complication of tags and tokens are ignored compared to Beautiful Soup and Selenium. Reason: It created a huge dependency on the website styling and HTML formats to define and parse for JSONify the whole raw data. In addition, we need to get access to the website which compiles all the data privacy policies. The API-based method requires the available API endpoints from the referrer and source, which will be accessed through a basic HTTP GET request with all the headers and parameters required to fetch the data.

A screenshot of a computer

Description automatically generatedIn this work, we applied mimic headers to access the website as the API endpoints contained some bugs. A total of 70K records were present, out of which 55% i.e., 40,000 data was scrapped for this project. The below figure illustrates the XMLHttpRequest page and the hierarchy of data is stored.

*Figure2: API endpoint response*

## Data Wrangling and EDA

The initial process of checking the dataset is carried out with all the necessary steps followed in the industry. The data is carefully examined and the fields are clearly explained to get the complete understanding of the data. The following representation gives overall column (field) details present in the data. A screenshot of a computer program

Description automatically generated

* Title: the title of a book
* id: an unique id for each book in the dataset
* price: the price of the particular book
* author: the author of the particular book
* publisher: the publisher of the book
* pub\_year: the year in which the book was published
* s\_title: the subtitle of the book
* edition\_num: the edition number of the book
* description: a detailed description of the book
* available: indicates the availability status, whether a book is in stock or not.
* sale\_date: the date of sale for the book
* short\_pub: sale date for the book in short (Jan 24)
* num\_of\_author: the number of authors associated with the book
* width: width dimension of the book (in cm)
* height: height dimension of the book (in cm)

After capturing what kind of data is present, the next step is to extract the meta information of the data. It includes but is not limited to the shape (total rows, columns), size (individual entries), data types and categories .Along with this, the missing values, duplicated rows and unique distinct values are calculated. Using pandas, the data is transformed and loaded into a dataframe for easy access and operations. The following table discusses the useful meta information of the data.

|  |  |  |
| --- | --- | --- |
| **S No** | **Analysis** | **Value** |
| 1 | Shape of the dataset(dataframe) | (39990 rows,15 columns ) |
| 2 | Dataset size | 598,500(approx.) ~ 4.5MB |
| 3 | Data types present in raw data | * Object(string) * Int64 * Float64 * Bool |
| 4 | Numerical and categorical(includes text) columns | * Numerical = 7 columns * Categorical =7 columns * Binary=1 column |

Using the metadata, the missing values are calculated across all the columns, total of 5 columns has a considerable amount of missing values present in it. After analyzing the impact and percentage of the missing data in the dataset.

A white background with black numbers and letters

Description automatically generated

*Figure3: missing values*

* “Subtitle” is not considered for any processing, as it is not having significance
* For the other columns like author, description and price where the missing value percentage is less than 14% is retained for further imputation or processing
* Experimenting on the edition number by creating two distinct data frames i) with imputed edition number ii) dropping the edition number column
* 'description' won’t be used for modelling, we consider keeping this column to use it for text data analysis.
* It is observed that there is no significant impact on removing the edition number or imputing it. As the values are less importance for the clustering technique (as per the basic clustering results with price, edition number)
* As per the analysis, there are around 9153 duplicated rows present in the dataset, as we closely observe,
* These are the books with no edition numbers but same authors
* The chances of the duplicates having different edition number is highly suspected
* The correlation analysis from the pandas profiling shows the weighted relation between various features
* The duplicate values which had no impact over the analysis are removed from the dataset for smooth analysis

A screenshot of a computer code

Description automatically generated

*Figure3: data inconsistency*

The given data has inconsistent data types present in it, for instance the column ‘sale\_date’ is of object (string) type. It is advisable to convert the datatypes of the columns which have the best feature set to support the clustering process. Data Types which are converted for processing are

* Sale\_date : converted from string to date time
* Availability : Boolean to numeric

**Feature engineering**

To support more of the clustering system, new features are included in the dataset. The following table demonstrate the feature columns created from the data

|  |  |
| --- | --- |
| **Original Column** | **New Feature** |
| Sale\_Date | Year, Month and Date are created   * Year and month are utilized for processing |
| Description | 1. Processed description 2. Key words (used for content based clustering) 3. Length of description |
| Price | Aggregated price (High ,Mid ,Low) |

Reason to create:

* Sale date is the crucial factor binding the relationship between the sales and the price of the book
* Price has 1307 unique values and processing this huge values might create some bias, so it is recommended to create three distinct categories like High, Mid, Low for easy computing
  + Insight
    - Most of the books price falls under 95 dollars, (75 percentile)
    - 95 percent of the books fall under the price of 270 dollars

A number and text on a white background

Description automatically generated

* Based upon the distribution of book prices, values between 0 - 60 is categorized as Low; 60 -270 as medium and more than 270 are high (costly)
* Price range of 60-80 is the most common as it is reasonable price to buy a book
* The proportionality of the book prices after categorising is displayed in the graph below.

A graph of a number of different sizes and numbers

Description automatically generated with medium confidence

* Key words are created by combining the description and the title of the book to generate book context or genre
  + NLTK module is utilized to process the text and get the key words
  + Usage of Key words are discussed further in the report.
* Finally the categorical columns are checked for unique values present and the values are observed to require some encoding technique later in the pre-processing techniques to have optimal results for our modelling.

## Natural Language Processing

As specified in the previous section, columns like description, title, author name, publisher name are text oriented. Processing of any text data requires a special and careful approach. During the pre-processing approach we conducted i)lowercasing the strings present in the text ii) removing punctuations iii) removing unwanted spaces iv) stopwords parsing v) spell checker.

A screenshot of a computer

Description automatically generatedThe pictorial representation is from pandas profiling, which will be discussed in detail along with the data visualization topic in this report. It shows the most frequent stopwords present in the description field.

* These stopwords are very often present in the book
  + Reason : Books are meant for everyone, so usage of simple and same words are very high compared to logical or technical contents
* Removal of other language words and the removal of emoji’s was done utilizing techniques from stackoverflow community guidelines.
* Regular expression is utilized with mapped emoji format which are standardized in the text format of the data.

Reason to process the text is to create a content based (little similar to LDA) clustering. It is achieved through cosine similarity between the two descriptive contents. It greatly helps in grouping the unlabeled data into particular groups. Term Frequency-Inverse Document Frequency is utilized to create vector representation of the text. The detailed process is covered in the later section of this document.

Visualization and pandas profiling:

The scraped data is visualized to check the distribution of data and find any interesting pattern using the pandas profiling method. Before checking the report using pandas profiling, the general histogram interpretation is done. It shows the spread of data points for each numeric column. We observed that most of the numeric data is skewed (both positive and negative). Later, the publication value counts are calculated and linked with the bar plot. It clearly shows data during 2014- 2018 is consider as the saturation point which ends the gradual increase of the publication over 3 decades. It might be due to various external factors and pandemic.

A graph of a number of years

Description automatically generated

The author-based publications are checked by plotting it by using the most frequent method. Here authors of various exam preparation and software testing manuals are listed with 40 publications under their name. The interesting part is that book data contains the word “book” as the most frequent word (14000 times) and it is not a surprise as this report might contain word “report” as the most frequent word.

A graph of red squares

Description automatically generatedA graph of a number of books

Description automatically generated

The word cloud describes the terms related to computer domain

A close-up of words

Description automatically generated

Using the pandas profiling report, complete details of the data points present in the dataset is inferred. It provides granular level details starting from spread-out rate to correlation analysis. The report is utilized in properly to make use of the stats and hidden information in the data with less time. A screenshot of a computer

Description automatically generated

The correlation between data points are observed from the pandas profiling A screenshot of a computer screen

Description automatically generated

## Encoding

Encoding is very handy when it comes to processing high dimensionality data features, it allows us to process the data with less impact on the originality of the data. In this activity, two distinct approaches are applied on the dataset

1. Encoding Categorical values.
2. Discretization of large numerical values to categories then to numeric representation.

When it comes to encoding of categorical values, the features considered are i) ‘affordability’ and 'Available'. Since there is no dimensionality concern, simple label encoder is used on it. Columns like year and price are very tedious to analyze as price is continuous value and the years are complex. Before encoding, the nature of both fields are analyzed. The price values are plotted to understand the segmentation logic. Years from the publication and sale are combined to get a clear picture on the distribution. The following table explains the encoding techniques applied on the dataset

|  |  |  |
| --- | --- | --- |
| **Original Column** | **Encoding type** |  |
| Affordability | Labelencoder |  |
| Available | Onehot encoding (dropped the collinear column) |  |
| pub\_year | 1. distinct values (experiment 1) 2. Discretization → Labelencoder  (experiment 2) |  |
| updated\_price | 1. distinct values (experiment 1) 2. Discretization → Labelencoder  (experiment 2) |  |

Observation:

* With ordinal encoding as our prime approach, the discretized value bins are encoded and tested for the clustering model in the analysis
* Both the encoding methods are experimented for better results in the clustering. The distinct values are time consuming process and the values are not standardized
* In the second approach, the ordinal values are encoded with standard deviation of 1
* We learned about pd.qcut on creating quantile based bins

## Outliers

Outliers are the true signal (data) which has value compared to noise, where it does not possess any value or significance. In this project we implemented various outlier detection strategies to catch the extreme values. The following table explains the various outliers detecting methods used,

|  |  |  |  |
| --- | --- | --- | --- |
| S No | Outlier Strategy | Description | Remarks |
| 1 | Prop Plot (Q-Q plot) | Statistical based approach to compare the data points with theoretical value measured. Refer the below figure | New learning for the team, to check for data points that deviate significantly from the reference line. Helping us to detect outliers. |
| 2 | IQR (interquartile range) | Comparing the value with lower and upper whisker data points of the numerical columns. | Simple and easy technique but not suitable for complex comparison and detection |
| 3 | Skewness and Kurtosis | Used to find the distribution symmetry of the data. It is used as validation step for the other outlier detection technique | Skewness: less than 1 → left skewed data vice versa for the other value (ideal value is 0)  Kurtosis: relativeness to the normal distribution  Value near to 3 is good fit |
| 4 | boxplot | Graphical representation of the data points using the same IQR approach | Easy to understand the outliers  Difficult to interpret the values on whole |

Observation:

* High right skewness for Price column distribution is definitely not normal.
* Most of the values fall outside the normal range by noticing the box plot for Price
* A screenshot of a computer code

  Description automatically generatedpub\_year shows left skewness and most of the values fall outside the lower bound
* other numerical variables are linear values, hence the linear lines on the chart

## Learning:

Probplot has a line called the reference line, which denotes the expected theoretical distribution. If the data points fall approximately along a straight line on the Q-Q plot, it indicates that the data follows the specified distribution.

A graph on a white sheet

Description automatically generated

* from the observation of unique values present in the current df\_5, publisher has 194 values --> this distinct will be difficult
* for encoding like one-hot as the dimensionality issue arises and creates zero in various columns (sparse matrix)
* Various combination of data points are plotted using pair plot and the potential outliers are displayed
* It is to be noted that the outliers are independent occurrence compared to noise where it brings some correlation factors
* From the pairplot, the updated price and the year columns states the combined outliers present in the data
* Distinct outliers are observed in author i.e., near 60 and publication year(near 1700)

A graph of a number of blue dots

Description automatically generated with medium confidence

## Handling outliers detection

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameters** | **Trimming** | **Flooring and capping** | **Logarithmic transform** |
| Column dropped | 54% (16913) | - | - |
| Distribution of data | **Even** but discrete as we dropped potential values | Smooth distribution | Normal distribution |
| Skewness(mean) | Negative | Near to zero | High Negative |
| kurtosis | Negative (less than 3) | Negative (less than 3) | Negative |
| Inference | Not recommended | Recommend for extreme edge cases | High recommended for time and big data |
| observation | Outliers are still present due to the diverse data points and median line | Outliers are total discarded | Outliers are scaled down and brought to even distribution |

* The outlier handling was interesting task to apply on the data and figure out the best suitable approach based on the nature of the dataset
* Here were tried to run with floor and capping method to find a best feature set
* Created different copies of data frame for the outlier handling step, each data frame are compared with other and the suitable one is taken into the pipeline. To reduce the complexity of the report, it is limited to the original state of the dataframe.

A graph with numbers and lines

Description automatically generatedA graph with a rectangle and a line

Description automatically generated

## Clustering

A graph with a line

Description automatically generatedIn this project we considered using two unsupervised models to understand the segregation of the data based on similarity and differences. K-means and DBSCAN is used here with **three different feature pair combinations** i) price and publication year ii) price and edition number iii)topic (description) and price.

**K-Means**

K-means uses the centroid-based clustering approach by initializing the K-value and clustering the similar data points together. Elbow method is used to find optimum K value for the given data by looking at the change in squared distance. The nose point where saturation starts is conifer to be near the optimum value. Here it is between 3 and 4. Later, the clustering model is applied on the data and the cluster classes are assigned. Scatter plot is used to easily interpret the clusters visually and sophisticatedly.

**A graph with different colored spots

Description automatically generatedPrice and publication year**

This combination gives clear distinct clusters. It clearly groups the data where the publication year is between early 80s to mid-21st century. The first clusters mapped along the extreme value we included to check the performance of the clustering. The second cluster is along the price below 150 and the last cluster contains the remaining data points. It clearly met the expectation of clustering analysis. Clustering is validated with silhouette score **K=3 🡪 0.74** , K=2 🡪 0.43. K=4🡪 0.46.

**Price and edition number**

A graph with numbers and dots

Description automatically generated

This combination does not provide any valid inference or cluster group in general. The edition number is not creating any impact on the created groups, the change in price creates the second distinct cluster and the best silhouette score formed is 0.47 which is not making sense to the clusters created with the previous combination.

**Description and price**

A graph of a clustering chart

Description automatically generated with medium confidence

Third combination of the clusters is formed by considering description length and price. This gives us slightly tightly packed clusters where the price determines the clusters over the increase in Y axis and the description length also changes clusters along the X-axis. The best silhouette score .57, which is worst compared to any other combinations as Kmeans are difficult to group on density based data.

**DBSCAN**

A graph showing the number of data

Description automatically generated**Price and publication year**

The density based clustering method applied here creates clear distinct clusters which are closer together and the distance data points are formed as separate clusters by increasing the price and year in the left side of the graph. The clustering visualization suggests that there has been a tendency for book prices to increase over time and that there are distinct groups or categories of books that can be identified based on price and publication year, with some books standing out as exceptions to the general pricing trends.

**Price and edition number**

A graph with numbers and colored dots

Description automatically generated

The density plot for this combination remains the same like K-means, illustrate that book prices correlate with edition numbers, with a visible trend of decreasing prices for books with higher edition numbers. It is not giving any useful information on the groups and it cannot be used for grouping

**Description and price**

A graph with dots and numbers

Description automatically generatedThis is the best plot so far for the density-based clustering method, as it represents the data that are closely packed and the remaining clusters are categorized into 2 different clusters for easy presentation.

**Inference and result:**

Out of all the combination and the clustering technology used in this project,

* K means clustering with the combination of price and year gives best possible clusters to the data
* K=3 and K=12 and K=15 is the optimum and good fit K value throughout the analysis
* DBSCAN gave a density plot accurately for the combination of description and price.
* The silhouette scores for the best K-means method give best score of 0.74 which validates our inference of best clustering approaches so far

In addition to this, the clustering technique is leveraged by applying NLP-based text vectorization and comparing it using a famous similarity method called cosine similarity.

## TF-IDF COSINE

* Compared the keywords in root Noun form with one another and the values are stored in a database
* Using K-means, the optimum cluster for this feature is found to be 15 and the whole data is fit to the clustering object

A screenshot of a calculator

Description automatically generated

It provides us a high dimensional graph, complex to carry out any interpret.

Later, T-SNE (t-distributed Stochastic Neighbor Embedding) is applied to the clustered points. It mapped the data points near to each other as one and the rest respectively.

A blue and green blotches

Description automatically generated

* Learning from this process is how to visualize a high dimensional points in graph without losing it originality.
* This is the near perfect, best fit clusters generated based on the GENRE of the books through the description. The genre of book is not our feature and yet we achieved this through NLP.
* Validation of cluster is manually done through comparing the description of all the books in one cluster and the results are more satisfactory.

**Conclusion and Future Work**

This project provides us clear understanding of how various clustering techniques are working and the principles behind the clustering. Further, the complete end-end cycle starting from scrapping the data and storing it for extended data cleansing process till the clustering validation process it gives how data processed in correct way provide useful information like genre which not extract buy yet created with the help of NLP and clustering technique . Furthermore, there are various area to improve our approach and steps. The data can be further processed with various combinations.

* Future work includes mapping the sale date with the price feature to extract the relation between how dates are impacting the sales of the book by clustering it by date wise price amount on each books.

# **References**

Prabhakaran, S. (2022, April 20). *Cosine Similarity – Understanding the math and how it works (with python codes)*. Machine Learning Plus. <https://www.machinelearningplus.com/nlp/cosine-similarity/>

Chandrimad. (2020, August 2). *Clustering with PCA, KMeans, Hierarchical, DBSCAN*. Kaggle. <https://www.kaggle.com/code/chandrimad31/clustering-with-pca-kmeans-hierarchical-dbscan>

Ssabrut. (n.d.). *GitHub - ssabrut/book-recommender-system: This project associated with my university for milestone project. A book recommender system using k-means clustering with content based approach from goodreads book dataset.* GitHub. <https://github.com/ssabrut/book-recommender-system>

Hasson, E. (2023, December 20). *What is scraping | About price & web scraping tools | Imperva*. Learning Center. <https://www.imperva.com/learn/application-security/web-scraping-attack/#:~:text=Web%20scraping%20is%20the%20process,replicate%20entire%20website%20content%20elsewhere>.

Rafaymemon. (2023, July 10). *OUTLIERS TREATMENT TUTORIAL BEGGINERS*. Kaggle. <https://www.kaggle.com/code/rafaymemon/outliers-treatment-tutorial-begginers>

Derksen, L. (2022, September 28). *Using T-SNE in Python to visualize High-Dimensional data sets*. Built In. <https://builtin.com/data-science/tsne-python>

## Dataset and Onedrive links

* Dataset - <https://mylambton-my.sharepoint.com/:x:/g/personal/c0910392_mylambton_ca/EcNOR8FqKpZIjLSAd6RMMgwBkL0XdUsWS2YNzf57-Jfc8g?e=Eq7kIx>
* pandas\_profiling- <https://mylambton-my.sharepoint.com/:f:/g/personal/c0910392_mylambton_ca/EkkJIqcSk9xAl8bWBSNWcngBioMC9-S8-n7jSKfmS6yIcg?e=nV9AfI>
* Whole working folder - <https://mylambton-my.sharepoint.com/:f:/g/personal/c0910392_mylambton_ca/ElTNVm15LMlLhwqns823Pn4BecucWfYN-5o-MjCpkLv6Zw?e=KDfTG0>
* Git hub - <https://github.com/svjai/AML_2203_project>

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated