Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Preprocessing and Clustering Techniques for Uncovering Demographic Insights: A Study of Book Preferences Using PCA And K-Means

Tuqa Muslim Yaqoob[1*], Enas Fadhil Abdullah[2]

[1]University of Kufa , Collage of Education , Dept. of Computer Science, Najaf, Iraq,tuqam.albatat@student.uokufa.edu.iq
[2] University of Kufa , Collage of Education for Girls, Dept. of Computer Science, Najaf, Iraq, inasf.alturky@uokufa.edu.iq
Corresponding author e-mail: tuqam.albatat@student.uokufa.edu.iq

https://doi.org/10.46649/fjiece.v4.1.1a.25.3.2025

Abstract. *Authorities have interest in the use of demographic data to understand users' behavior due to its importance in making industries unique. In the analysis of the demographic data regarding books' preferences this study makes a use of the Principle Component Analysis (PCA) and K-means clustering analysis. First a PCA was used in an attempt to reduce the dimensionality of the given dataset, the goal was to maximize the variance retained in the data while at the same minimizing the raw data complexity to 95%. Last, in an effort to classify the data into different demographic areas by age as well as geographical region, the K-means clustering was applied. The outcomes implied in the study show that PCA, as well as K-means clustering, are efficient in distilling information from extensive and great-detail demographic data. The insight arising from the results is additional information in the area of reader preference, as well as suggestions for future marketing and recommendation approaches. These methods were used on the book dataset that contain user and book data, and the results proved that every demographic had preferences with different books. On top of knowledge derived from analysis of the results, it offers further understanding of the readers' choice of content and indicates areas where marketing and recommendation strategies may be headed in the future. As a result of clustering customers, businesses can for example, target the marketing campaigns and offer appropriate books to the specific ages and geographical areas.*

Keywords: *K-Means Clustering, Dimensionality Reduction, Principal Component Analysis, Demographic Data, Book.*

1. INTRODUCTION

With competition being tight in the present world, business marketers need to assess consumer behavior in a bid to be in a position to fulfill the objective of delivering satisfying experiences to the consumers. Similarly to most other industries, data analytics in the book industry mostly serves to gain insights into customers' behavior and trends. Interestingly, demographic information is one of the most helpful for exploring the details of the readership's activity and, consequently, for addressing the polyphony of the market. Consumers are today becoming the benchmarks within the book industry, primarily through use of data analytics to identify consumer preferences and trends. Information concerning demographics especially proves helpful in understanding various parameters of reading behaviours thus helping businesses address the needs of various demographic segments. In this study, machine learning techniques such as PCA and K-means clustering will be used to analyze demographic characteristics that affect book preferences. The primary objective is to identify and categorize such fields

of the user's profile as age and/or location. It is in that sense, by running PCA on the data and then categorizing it with K-means, that one aims for insights which would be useful in practice to get a sense of how people categorized by their demographic profile interact with books. The future outcomes of such works are apparently valuable for the formation of recommendation models, the formation and design of marketing concepts and m-commerce experiences, and for overall improvement of applications and their user interface effectiveness.

It intends to demonstrate how PCA & K-Means clustering can be employed, in order to evaluate the discovered demographics for delivering efficient and effective insights to the organization or researchers who want to target and understand a variety of reader types.

2. RELATED WORK

This paper examines the suitability of applying two well-known clustering techniques namely Principal Component Analysis (PCA) and K-means clustering in the domains of machine learning and data analysis. A large number of studies provide evidence that PCA is useful in the exploration of a large number of dimensions of a complex dataset while preserving the large variance. For instance, Jolliffe [28] proposed the applicability of PCA in a vast field of application such as image processing and genomics. Similarly, different resources indicate that K-means clustering is an easy method and useful for dividing data into meaningful clusters. Since its original development by MacQueen [29], K-means has since been applied in a variety of context including market basket analysis, image compression, and customer segmentation.

Some of these method have been used in this research to establish patterns and trends relating to consumer behaviour as well as, demographic data. For example, in the study by Smith, et al, [30] consumer purchasing behaviour was analysed with the help of both PCA and K-means, and valuableSo, for example, a study by Smith et al. [30] identified important prospects regarding PCA and K-means with reference to certain customer segments by analysing consumer purchasing behaviour. Using these techniques on health care data, Johnson and Miller [31] conducted another relevant study which found out demographic segments of significant health concern.

In this research, building on the prior PCA and clustering analysis, the current study employs a dataset of book interests and applies PCA and K-means clustering. This research seeks to provide real life recommendations which, when adopted would enhance social commerce and recommendation system for the book industry specifically focusing on the age and geographic location data. When these methods are used collectively, there is a clear conceptual foundation of demographic knowledge to support business dynamics.

3. MACHINE LEARNING TECHNIQUES

Nothing is more exciting to have ML as an inevitable foundation amongst the technology where it revolutionizes many fields and the preparations for the progressive advancement. The form of this sort AI is machine learning; it provides the systems the ability to learn from the input values and make decisions based on them independently of a human input. This makes machine learning important since they are

capable of handling big data, making accurate predictions and are also capable of solving some tasks…effectiveness and efficacy.

Clustering is an important area of machine learning. What the clustering algorithm does is partition data points into clusters that are as similar as possible. A data mining process must uncover subtleties that are invisible to the human eye. Data clustering provides a unique method of sorting big data, enhancing cognition of abstract patterns, and enabling more intelligent decision-making. By identifying these clusters, machine learning can analyze data more effectively, potentially leading to improved results. With the progression of machine learning on the rise, the importance of clustering and the other techniques in the data analysis will progressively increase the pace and scale up the innovation and transformation in different sectors. Rather, the widespread use of clustering and machine learning falls into various categories, including boosting analytic value, improving data analysis, defining action strategies, refining assessment, and enriching decision-making. Knowledge of the above techniques is essential, as they open up great possibilities of growth and transformation in today's world of data control [4-6].

In general, there are three categories into which machine learning falls: The three main learning categories are supervised learning, unsupervised learning, and semi-supervised learning. Both categories use different methods and are suitable for different tasks and problems [7, 8].

A) Supervised Learning

Supervised learning involves learning with examples that are associated with a particular output signal that the model is to learn [11]. This allows the model to understand or be programmed to predict outputs in relation to the inputs as held in the input-output pairs. This is usually utilized in activities such as classification and regression. With classification, the model finds and recognizes category labels, while in regression, the model approximates on a continuum [12]. All over the world, diverse areas such as image and voice recognition, spam control, and disease diagnosis use supervised learning.

B) Unsupervised Learning

Unsupervised learning, on the other hand, does away with the labelled data in as much as it deals with the unlabelled data. During the training, the model never gets explicit outputs or labels, as is the case with other models [11]. Rather, it aims at finding some structure that lies implicitly within the data [12]. This type of learning is highly effective for EDA and other data analysis tasks that require independent discovery of structure in the data. Unsupervised learning commonly employs the following techniques:

- **K-means:** This algorithm divides the data into k clusters, assigning each data point to the cluster whose mean is closest[13].
- **Hierarchical Clustering:** This algorithm creates a hierarchy of clusters by either merging smaller clusters into larger ones (agglomerative) or dividing larger clusters into smaller ones (divisive).[14].
- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN**): This algorithm makes the formation of clusters through density reach and makes it possible to identify clusters of all shapes and sizes as well as handle noise effectively [15].

The strength of unsupervised learning lies in its ability to uncover hidden patterns and associations that manual analysis of structured datasets would likely miss. Most commonly, it is useful in situations where the labelled data is scarce or expensive to come by.

C) Semi-Supervised Learning

Semi-supervised learning is in-between supervised and unsupervised learning. In fact, it works with a small amount of labelled data bundled with a significantly larger number of unlabelled data during the training stage. This idea can significantly improve learning accuracy because it is often difficult to capture a dataset with the correct labels. If some of the data is labelled while others are not, then using both results in the best technique known as semi-supervised learning, which has been proven to perform better than purely supervised or purely unsupervised learning in areas such as text classification and diagnosis of diseases through images, among others [16].

4. DIMENSIONALITY REDUCTION

PCA is a multivariate method that was designed to work with data tables with observations that are described by many related quantitative dependent variables. The principle and principal aim of PCA is to analyse the original data and replace the original multidimensional data with a set of linearly orthogonal variables referred to as the principal components [17, 18].

The core concept of PCA is dimensionality reduction. When working with a dataset that includes many interrelated variables, PCA aims to reduce the dimensionality while preserving as much of the dataset's variation as possible. The process accomplishes this by transforming the data into a new set of uncorrelated variables, known as principal components. These components are ordered so that the first few retain most of the variation present in all the original variables [19, 20].

PCA is arguably the fastest and most commonly used matrix factorization approach. In the category of unsupervised techniques called matrix factorization, there is a set of principled ways to show the low-dimensional structure of data while keeping as much information from the original set as possible [21].

In PCA, each principal component is a linear combination of the original features with optimal weights. The first principal component captures most of the variation in the data, and each subsequent component captures the maximum remaining variation [22]. Thus, PCA is a powerful tool for reducing a large number of dimensions to a much smaller, manageable number, making it highly useful for data analysis and interpretation.

PCA helps in [19]:

- **Enhanced Visualization:** High-dimensional datasets can be difficult to visualize and interpret. Dimensionality reduction techniques project these datasets into lower dimensions (such as 2D or 3D), facilitating better visualization and understanding.
- **Improved Model Performance:** By eliminating irrelevant or redundant features, dimensionality reduction enhances the accuracy and efficiency of machine learning models.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

- **Reduced Overfitting:** Simplifying datasets by reducing the number of features helps decrease the risk of overfitting, ensuring that models generalize better to new, unseen data.
- **Noise Reduction:** Removing less informative features reduces noise in the data, resulting in more robust models. Figure (1) shows the steps of PCA technique[23].

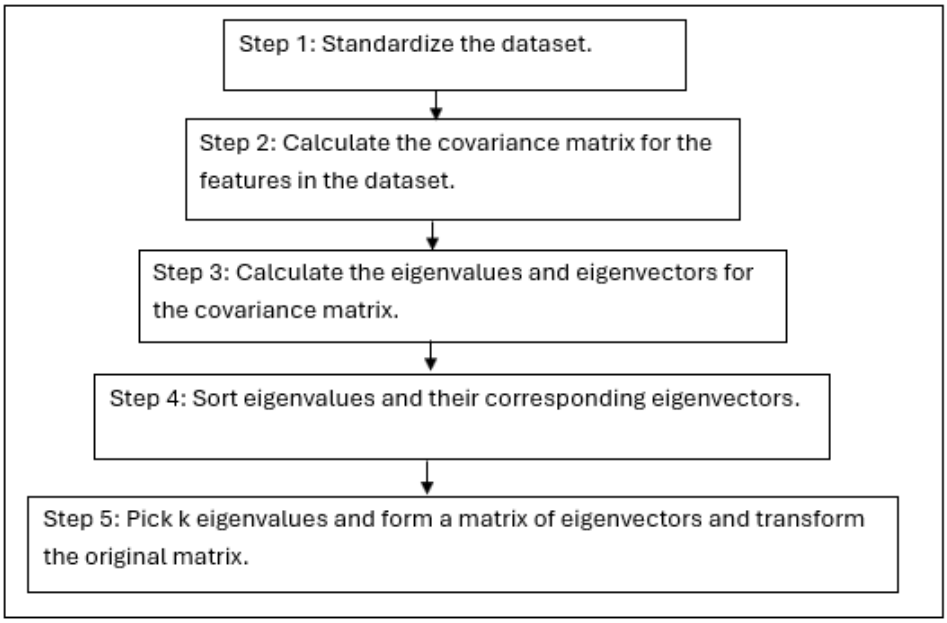


Fig.1.The steps of PCA.

5. WORK TECHNIQUES

Figure (2) explained methodology of work, which consists of three stages, the first stage is preprocessing. Second stage is dimensionality reduction and finally community detection.

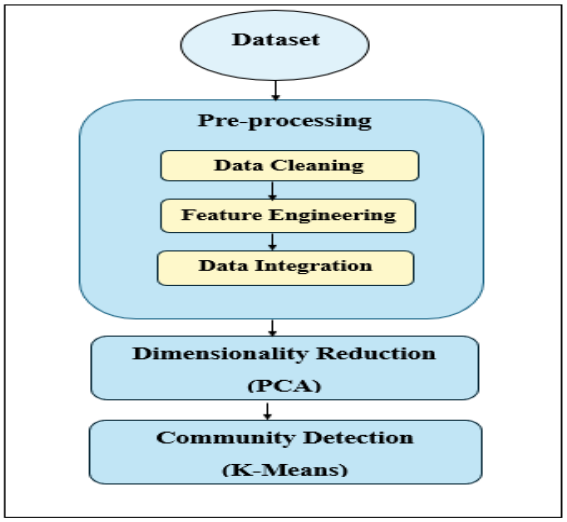


Fig.2. Methodology of work.

5.1. PRE-PROCESSING

An essential step in the process is pre-processing, and the following outlines the steps involved.

- ***Data Cleaning***

1. *Handling Missing Values:*

    - **Users Table (Age Column):** Missing ages were imputed using the mean age, which provides a straightforward and effective way to estimate the missing data.

    - **Books and Ratings Tables:** Given the minimal missing data, entries with missing values were removed to avoid any significant impact on the analysis.

2. *Removing Duplicates:*

    Using SQL, it has been recognized and removed with the focus on the frequent BookID and UserID joined pairs. This was necessary in order that each record was distinct and no distortion of data occurred.

3. *Handling Inconsistent Data:*

    Standardizing Age Values: Some limitations included People often give their age in an elastic manner hence restricted age to between 5 to 95 years since we were to work with real data.

- ***Feature Engineering***
    - **Age Binning:** The age data was capped and binned into categories that makes analysis easier, reduces the effects of extreme values and improves on the visualization of the results.

    - **Location Normalization:** Corrected typographical errors in spelling, and standardized geographic coordinates to contain only country level data. This approach helped big time in ensuring a consistency and accuracy on the demographic representations.

- ***Data Integration:***

    We merged the users, books, and ratings tables together for analysis, creating key attributes based on User ID. This added advantage for integration was that it provided a comprehensive view of the users' relationships and choices, thereby enhancing the data evaluation.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

The following algorithm outlines the pre-processing steps taken to ensure the dataset's quality for analysis:

Algorithm (1): Pre-processing on Book Dataset (PoBD)

Input: Dataset (user, book, and rating files)
Output: Pre-processed Unified Dataset
Begin
Step1: Data Cleaning
- Handle Missing Values:
  - Fill missing ages with the mean.
  - Remove minimal missing values in Books and Ratings tables.
- Remove Duplicates:
  - Identify and delete duplicate entries.
- Correct Unformatted Data:
  - Ensure ages are between 5 and 95 years.

Step2: Feature Engineering
- Age Binning:
  - Categorize age into groups: Child, Teenager, Young, Senior, Super Senior.
- Location Normalization:
  - Correct spelling errors in location names.
  - Standardize entries to use only country names.

Step3: Data Integration
- Merge User, Book, and Rating tables using User ID as the key.

Step4: Prepare Data for Analysis
- Verify data format and readiness.

Return: Pre-processed Unified Dataset
End.

5.2. DIMENSIONALITY REDUCTION USING PCA

Subsequently, after cleaning the data, we transformed the data using the Principal Component Analysis (PCA) algorithm to do dimensionality reduction, keeping as much variability in the new dataset as possible.

The process involved several key steps: First, we preprocessed categorical and textual data into numerical form using the right encoding functions since it is required to qualitatively analyze numerical data to include categorical data and texts into the PCA. Subsequently, in order to reduce the impact of features that could excite PCA, we normalized the data set so as to maintain optimal variances of the initial features. This was further decided according to the explained variance ratio as it tries to capture the

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

initial number of components that accounts for at least 95% of the total variance as in figure(4), thereby eliminating noisy and redundant features. Subsequently, we performed the PCA on the given dataset to apply transformation on the dataset into a new matrix that consists of principal components only. Furthermore, heatmaps were employed to display the correlation matrix as in figure (3), as well as the relations between the main components to map out the contribution of every component to decide on the variance of the data and detect patterns that would not be visible in the raw data.

Lastly, they discover that preprocessing the high dimensionality of the data using techniques like principal component analysis (PCA) reduces noise and dimensionality and subsequently improves the subsequent machine learning models like clustering algorithms like K-means. In addition, PCA improves interpretability by reducing data into principal components, making work on the results structure easier in terms of determining factors that cause variability in demographic information.

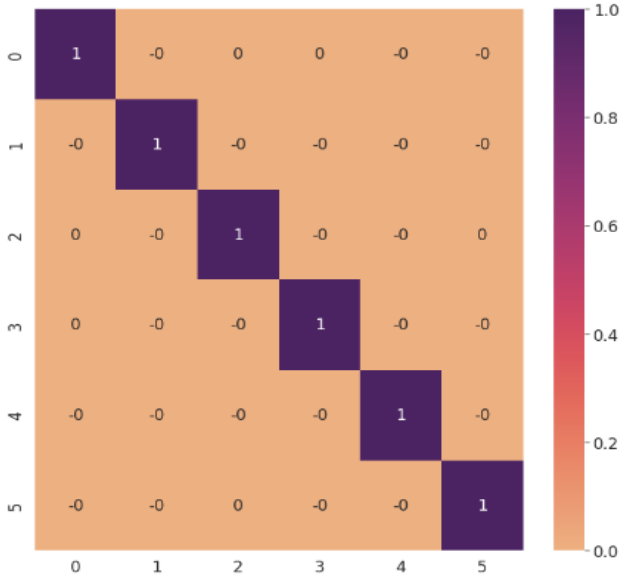


Fig.3. Heatmaps.

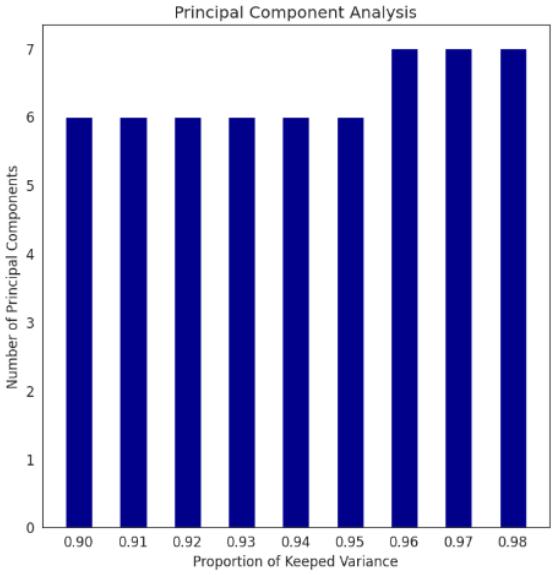


Fig.4. Proportion of Variance.

5.3. K-MEANS CLUSTERING ON REDUCED DIMENSIONS

Next, since the PCA dimensionality reduction algorithm was used to identify higher variance dimensions, K-Means clustering was used to group the book's data. This method also centered its work on major features derived from the principal components to enhance the clustering. Below is a summary of the K-Means clustering application on the reduced dimensions:

- **Selecting the Number of Clusters (K):**

  Elbow Method: To decide the number of clusters it was decided to use the elbow method. Here, the WCSS is graphed against the different numbers of clusters, K, and the position called the 'elbow point', where the rate at which the WCSS is declining starts to reduce, is found [11]. As for the book's dataset, it was found that the elbow peaked at 14 as in figure(5), so that number of clusters is sufficient to capture the structure of the data set without overcompensating for it.

- **Applying K-Means on Reduced Dimensions:**

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

The K-Means algorithm was used to divide the dataset into 14 clusters using the principal components that were found through PCA. By concentrating on the most important patterns that PCA had identified, we were able to effectively group data points with comparable attributes.

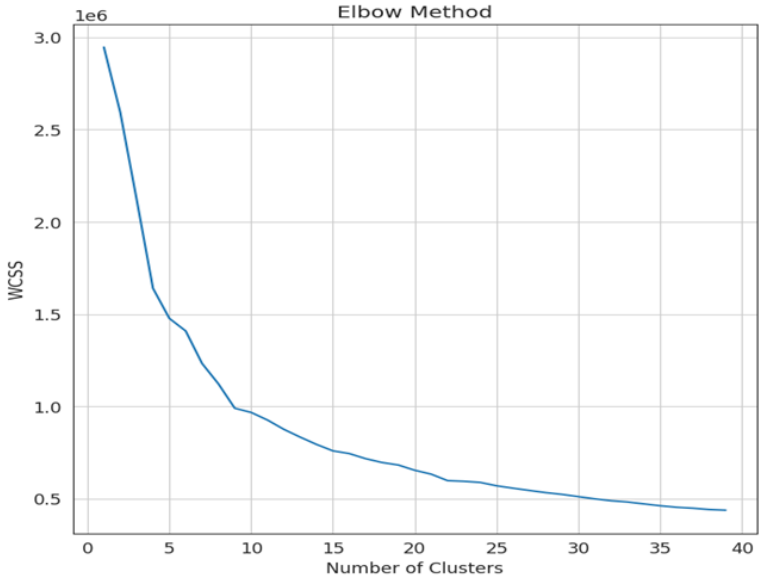


Fig.5. Elbow Method.

6. RESULTS AND DISCUSSION

6.1. RESULTS

There is a dataset of books available on Kaggle [24] where this research source's raw data is extensive. This dataset's various and numerous features make it suitable for use in recommendation systems, clustering, and sentiment analysis. It consists of three distinct tables, each providing specific information about books:

♦Books Table: It is as follows: Each book's metadata primarily includes the following: the book's title, the author, and the year of publication.

♦Users Table: The following table provides details of users in terms of age, location, and any other relevant information for the ten users in consideration. This information is helpful for the company to understand the users and the division of the target market.

♦Ratings Table: This table contains the user's opinions on books and contains fields like the user identifiers, book identifiers, and the rating. We use it for both analyzing user requirements and manufacturing recommendation models.

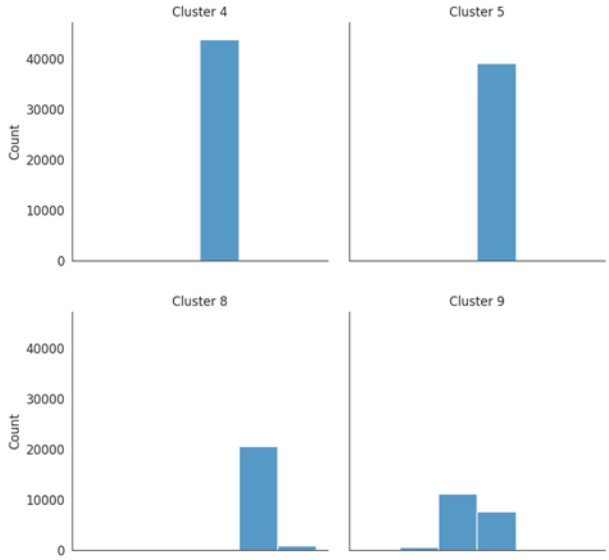
The related identities of all three tables enable the examination of information in multiple dimensions and the coordination of various characteristics.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Table 1. Contents of dataset

Table	Contents
BT_ table	ISBN,Title,Author,Year,Publisher and Image
RT_ table	ISBN,User_ID and Rating
UT _ table	User_ID,Location and Age

The 14 clusters are rather helpful for gaining a better understanding of more distinct categories of users, as well as their propensity to read depending on certain demographic characteristics such as age and geographical location. For example, young readers from urban areas of the sample were more likely to cluster than the older readers. Figure (6,7,8) shows information for each group.



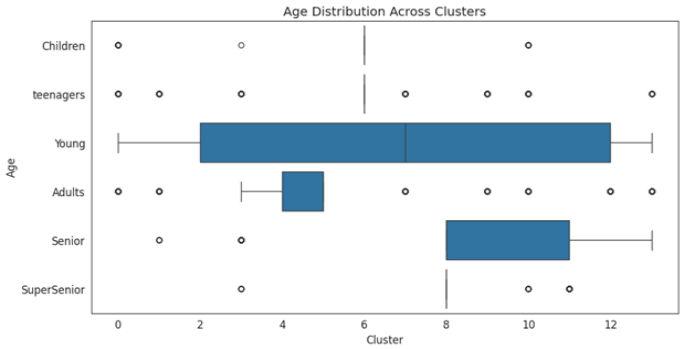


Fig.6. Demographic Information (Age and Country) Across Clusters. **Fig.7. Age Distribution Across Clusters.**

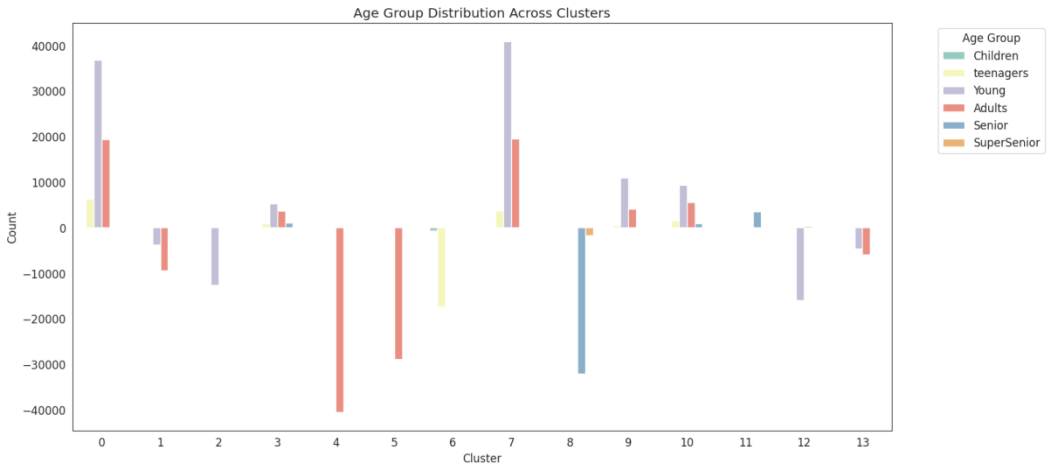


Fig.8. Age Group Distribution Across Clusters.

6.2. DISCUSSION

These clustering results are especially useful to recommendation and personal marketing systems. This is to enable business set their sight on the right demographic segment because of the characteristics that clustering sectorsizes. For instance, a publisher may decide on specific key clusters as Young Adults and ensure that the most advertisements targeting Young Adult novels are directed there.

These ideas can also be used to develop contents. The situation allows authors and publishers to search for trends within each cluster and respond to them with producing new books, which, in turn, are tailored for meeting the desires and orientations of the target audiences. A publisher may need to invest more money either obtaining or advertising historical fiction titles if a definite cluster has an interest in that particular type of literature.

As mentioned these clustering findings are very useful in a personalised recommendation system. That way, platforms using demographic information can better recommend books, thus creating more user satisfaction and interaction. For example, let a recommendation engine would understand a user would prefer a set of books and make recommendations for a user, more broadly by recommending books, which is popular among the users at the user's age or at the region of the user.

Taking everything into consideration, the use of PCA and K-means clustering on demographic data provides firms in the book industry robust instruments. From it, one can get a better understanding of what readers want in more depth, manage marketing and promotional techniques more effectively, and improve the development of specific recommendation services. By obtaining these, businesses shall be able to satisfy it customers and gain their loyalty hence promoting growth and success for operations in a competitive economy.

7. CONCLUSIONS

Demographic analysis of the book dataset was made possible by the use of K-Means clustering applied to data with PCA reduction. According to this analysis, both country and age had an impact on reading habits. Using the elbow method, we selected 14 clusters to simultaneously capture complex patterns without overcomplicating the analysis. Incubating decisions regarding solutions and techniques for trying them out thus showed that a clustering of different user segments in the form of users of books could provide a different engagement. Notably, the results also identified the favorite books for each group with average rating as in appendix (A, B), which would further the knowledge about various reading interests. Such findings enable organizations to enhance the efficiency of the recommendation together with the personalized marketing systems in delivering satisfying solutions to their clients.

In addition, this approach provides a foundation for subsequent studies and articles related to recommendation systems. If different demographic patterns are pinpointed clearly and elaborately, then researchers and developers can propose algorithms of higher sophistication, which, as a result, can provide apt recommendations to users, thus helping users yield the anticipated satisfaction levels.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Cluster 1:

	Book Title	Average Rating
8	1,911 Best Things Anybody Ever Said	10.0
9	100 Malicious Little Mysteries	10.0
10	1001 Things Everyone Should Know About the Universe	10.0
11	101 Things You Don't Know About Science and No One Else Does Either	10.0
12	101 Ways to Make Your Child Feel Special	10.0
13	18 Best Stories by Edgar Allan Poe	10.0
14	300 Incredible Things for Travelers on the Internet	10.0
15	300 Incredible Things to Do on the Internet -- Vol. I	10.0

Appendix A: Cluster 1, favorite books.

Cluster 11:

	Book Title	Average Rating
88	1,000 Years, 1,000 People: Ranking the Men and Women Who Shaped the Millennium	10.0
89	100 Selected Stories (Wordsworth Classics)	10.0
90	101 Dalmatians	10.0
91	20,001 Names For Baby : Revised and Updated	10.0
92	30 Days of Night	10.0
93	300	10.0
94	7 Kinds of Smart: Identifying and Developing Your Multiple Intelligences	10.0
95	A CHRISTMAS CAROL	10.0

Appendix B: Cluster 11, favorite books.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science (80-.).*, vol. 349, no. 6245, pp. 255–260, 2015, doi: 10.1126/science.aaa8415.

[2] G. W. Milligan and M. C. Cooper, “Methodology Review: Clustering Methods,” *Appl. Psychol. Meas.*, vol. 11, no. 4, pp. 329–354, 1987, doi: 10.1177/014662168701100401.

[3] J. Crawford, J. Gower, J. Lingoes, W. Rhee, F. J. Rohlf, and W. Sarle, “ An examination of procedures for determining the number of clusters in a data set ,” vol. 50, no. 2, pp. 159–179, 1985.

[4] Hamad SM, Al-bakri AA. "Enhancement of Brain Computer Interface System Based on Artificial Intelligent Technique". Al-Furat Journal of Innovations in Electronics and Computer Engineering. .vol. 3, no. 1, pp. 15–25. doi.org/10.46649/fjiece.v3.1.2a.11.4.2024.

[5] Younis HA, Ruhaiyem NIR, Ghaban W, Gazem NA, Nasser M. A Systematic Literature Review on the Applications of Robots and Natural Language Processing in Education. Electronics. 2023; 12(13):2864. https://doi.org/10.3390/electronics12132864.

[6] Hayder, I.M.; Al-Amiedy, T.A.; Ghaban, W.; Saeed, F.; Nasser, et al " An Intelligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Ad-vanced Alert System". Processes 2023, vol. 11, no. 2, 481.doi.org/10.3390/pr11020481

[7] Msallam MM. "An approach to hide an audio file in image using LSB technique". Al-Furat Journal of Innovations in Electronics and Computer Engineering.vol. 2, no. 2, pp. 1–7, 2023.

[8] Hayder, I.M.; Al Ali, G.A.N.; Younis, H.A. "Predicting reaction based on customer’s transaction using machine learning approaches".Int. J. Electr. Comput. Eng. 2023, 13,pp. 1086–1096. doi.org/10.11591/ijece.v13i1.pp1086-1096.

[9] L. Cao, "Data science: A comprehensive overview," ACM Comput. Surv., vol. 50, no. 3, 2017, doi: 10.1145/3076253.
[10] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, 2021, doi: 10.1007/s42979-021-00557-0.
[11] R. K. Cersonsky and S. De, "Unsupervised learning," *Quantum Chem. Age Mach. Learn.*, pp. 153–181, 2022, doi: 10.1016/B978-0-323-90049-2.00025-1.
[12] I. H. Sarker, A. S. M. Kayes, and P. Watters, "Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0219-y.
[13] P. Li, Y. gen Ding, P. peng Yao, K. min Xue, and C. ming Li, "Study on High-Temperature Flow Behavior and Substructure and Texture Evolution of TA15 Titanium Alloy," *J. Mater. Eng. Perform.*, vol. 25, no. 8, pp. 3439–3447, 2016, doi: 10.1007/s11665-016-2173-6.
[14] D. Jones and W. D. Grant, "Peter Henry Andrews Sneath," *Biogr. Memiors Fellows Rpyal Soc.*, vol. 357, no. November 1923, pp. 337–357, 2013.
[15] M. Daszykowski and B. Walczak, "2.26 - Density-Based Clustering Methods," *Compr. Chemom. Chem. Biochem. Data Anal. Second Ed. Four Vol. Set*, vol. 2, pp. 565–580, 2020, doi: 10.1016/B978-0-444-64165-6.03005-6.
[16] Z. Song, X. Yang, Z. Xu, and I. King, "Graph-Based Semi-Supervised Learning: A Comprehensive Review," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 11, pp. 8174–8194, 2023, doi: 10.1109/TNNLS.2022.3155478.
[17] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010, doi: 10.1002/wics.101.
[18] Younis, H. A., Ruhaiyem, N. I. R., Badr, A. A., Eisa, et al ." Creating the Hu-Int dataset: A comprehensive Arabic speech dataset for gender detection and age estimation of Arab celebrities". Biomedical Signal Processing and Control, 96, 106511.
[19] I. T. Jolliffe, "Principal components," *Data Handl. Sci. Technol.*, vol. 20, no. PART A, pp. 519–556, 1998, doi: 10.1016/S0922-3487(97)80047-0.
[20] Pearson K., "Pearson, K. 1901. On lines and planes of closest fit to systems of points in space.," *Philos. Mag.*, vol. 2, pp. 559–572, 1901.
[21] L. L. Hsu and A. C. Culhane, "Impact of Data Preprocessing on Integrative Matrix Factorization of Single Cell Data," *Front. Oncol.*, vol. 10, no. June, 2020, doi: 10.3389/fonc.2020.00973.
[22] S. Deegalla, *Nearest Neighbor Classification in High Dimensions*, no. 24. 2024. [Online]. Available: http://urn.kb.se/
[23] J. Shlens, "A Tutorial on Principal Component Analysis," 2005.
[24] Book dataset (kaggle.com)
[25] D. Kim and S. Kim, "Comparing patterns of component loadings : Principal Component Analysis (PCA) versus Independent Component Analysis (ICA) in analyzing multivariate non-normal data," 2012, doi: 10.3758/s13428-012-0193-1.
[26] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.
[27] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 975–8887, 2014.
[28] I. T. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics, 2002.

[29] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 1967.
[30] J. Smith, A. Brown, and K. Lee, "Analyzing Consumer Purchasing Behavior Using PCA and K-means Clustering," Journal of Business Analytics, vol. 12, no. 3, pp. 45-60, 2018.
[31] R. Johnson and S. Miller, "Demographic Patterns in Healthcare: Insights from PCA and K-means Clustering," Journal of Health Informatics, vol. 15, no. 2, pp. 89-104, 2020.