University of Essex

Online

# Transcript

**Link to the video: here**

Hello! Good morning from Mexico! My name is Daisy, and this is my research proposal presentation. The intention is to fulfil the learning outcomes by critically evaluating existing literature, research design, and methodology for my chosen topic.

## Contents

For the content section, I will follow the guidelines established for research proposals by the Computing Department at the University of Essex Online. I will begin by introducing the project title, followed by an explanation of the significance of my research. I will then present the main research question that guides my study and outline the primary aims and specific objectives I aim to achieve.

Next, I will summarize the key literature that informs my research and describe the methodology and research design I will employ. I will also address any ethical considerations and risk assessments associated with my research. Additionally, I will provide a description of the artefacts that will be created. And finally, I will present a timeline of the proposed activities.

## 1. Project Title

My project title is "**Evaluating K-means Clustering for Customer Profiling in B2B Ecommerce**". I chose this topic based on the extensive research and feedback I received during my literature review. This investigation highlighted the potential of K-means clustering to improve customer in the B2B ecommerce sector.

## 2. Significance

- **Importance of Customer Profiling:** B2B (**business-to-business**) ecommerce transactions are characterized by fewer but more diverse customers compared to B2C (**business-to-consumer**), contexts.
- **Challenges in B2B Customer Segmentation:** Customer profiling in B2B ecommerce is necessary for personalizing business efforts, improving customer engagement, or driving sales. A customer-centric approach can foster loyalty, facilitate cross-selling, and ultimately increase revenue.
- **Potential of K-means Clustering:** It is a data-driven methodology, offers promise in segmenting B2B customers based on purchasing behaviours and demographic characteristics. However, its application is mostly limited by the complexity and scale of B2B datasets.
- **Theoretical and Practical Contributions:** The findings will contribute to the theoretical understanding of clustering algorithms and provide practical insights for businesses.

## 3. Research Question

The central research question guiding my study is:

*How effective is K-means clustering in segmenting B2B ecommerce customers based on purchasing behaviours and demographic characteristics?*

This question aims to explore the efficacy of K-means clustering as a tool for identifying distinct customer segments, thereby enabling more precise and effective marketing efforts.

## 4. Aims and Objectives

The primary aim of this research is to evaluate the effectiveness of K-means clustering for customer profiling in B2B ecommerce. By critically analysing the potential and challenges of this algorithm, the study seeks to develop our understanding of its application in segmenting B2B customers based on purchasing behaviours and demographic characteristics.

The objectives include identifying the strengths and limitations of K-means clustering in handling complex B2B datasets, exploring how it can improve targeted marketing strategies, and proposing recommendations for optimising its use in the B2B ecommerce context.

## 5. Key literature

Leveraging my experience in marketing and familiarity with the topic, literature selection commenced over January 2024. The search focused on recognized academic databases and search engines, such as ACM Digital Library, IEEE Xplore Digital Library, the Essex Online Library, ScienceDirect and Google Scholar. Search terms were tailored to involve key concepts, and filters were applied to refine search results, such as credibility. The key literature used were:

**Azevedo and Gartner (2020): Market Segmentation in Credit Markets.**

Discusses market concentration and competition dynamics.

Offers valuable parallels for understanding customer segmentation in B2B ecommerce through insights into market dynamics.

**Borlea et al. (2021): Unified Fuzzy C-means and K-means Algorithms.**

Proposes refinements to clustering techniques.

Highlights advancements in clustering methods, suggesting potential improvements for K-means in B2B contexts.

**Essayem et al. (2022): RFM Features and K-means Clustering.**

Explores customer clustering based on Recency, Frequency, and Monetary (RFM) features.

Demonstrates practical applications of K-means clustering in segmenting B2B customers using RFM data.

**Mussabayev et al. (2022): Scalability of K-means for Big Data Clustering.**

Discusses scalability and efficiency considerations.

Critical for businesses dealing with large and dynamic B2B datasets, ensuring K-means can handle the scale.

**Nouraei et al. (2022): Comparative Analysis of Clustering Algorithms.**

Compares various clustering techniques for heterogeneous datasets.

Provides insights into the strengths and weaknesses of K-means relative to other clustering methods.

**Zare and Emadi (2020): Customer Satisfaction Using Improved K-means.**

Focuses on customer satisfaction analysis using an improving K-means algorithm.

Highlights the role of clustering techniques in understanding customer preferences and satisfaction in B2B ecommerce.

## 6. Methodology

The main points for my methodology are:

**Research Design:** To achieve a comprehensive understanding of B2B customer segmentation, I will employ a mixed-methods approach, combining quantitative data analysis with qualitative insights (Xiaoyu, 2023).

**Data Collection:** The data collection process involves utilizing a dataset provided by the company where I work, after obtaining the necessary permissions. This dataset includes transactional and demographic information of B2B customers.

**Data Preprocessing:** To ensure uniformity and comparability, I will normalize the data and handle missing values using imputation techniques.

**Feature Selection**: Key features for clustering will be selected based on criteria such as Recency, Frequency, Monetary value (RFM), and Tenure. Additionally, domain knowledge will be applied to include relevant features specific to B2B transactions.

**Clustering Process:** The clustering process begins with determining the optimal number of clusters using methods like the Elbow Method and Silhouette Score. The K-means clustering algorithm will be applied iteratively, assigning data points to clusters and recalculating cluster centroids until convergence is achieved.

**Validation and Evaluation:** Cluster quality will be assessed using internal validation metrics such as Silhouette Score and Davies-Bouldin Index. External validation will involve qualitative feedback from business stakeholders.

**Scalability and Efficiency Considerations:** To handle large-scale B2B datasets, big data clustering techniques will be implemented. The clustering process will be optimized for scalability and efficiency.

**Data Analysis**: The analysis of segments will involve examining the characteristics and behaviors of each customer segment. Patterns and trends within the segments will be identified.

**Tools and Software:** Python will be utilized for implementing K-means clustering. Power BI will be employed to present clustering results and insights effectively.

## 7. Ethical considerations and risk

In conducting this research, I will address several key ethical considerations and risk assessments.

**For the Ethical Considerations:**

**Data Privacy and Confidentiality**: To protect personal and sensitive information, all customer data will be anonymized. I will obtain explicit consent from the company to use the dataset, ensuring compliance with relevant data protection laws and regulations, such as GDPR.

**Informed Consent**: I will seek informed consent from all stakeholders involved in the data provision and analysis process.

**Data Security**: Robust data security measures will be implemented to prevent unauthorized access or data breaches. This includes utilizing secure storage solutions and encryption methods for handling and storing data.

**Transparency and Accountability**: Transparency will be maintained throughout the research process by keeping detailed records of data collection, preprocessing, analysis, and validation procedures.

**Risk Assessment**

**Potential Risks**: Potential risks include data breaches, bias in data analysis, and misinterpretation of results. Unauthorized access to sensitive customer data could lead to privacy violations and reputational damage. Biases in the data or clustering process could result in inaccurate segmentation and conclusions, while misinterpretation of the clustering results could lead to ineffective business strategies.

**Mitigation Strategies**: To mitigate these risks, I will implement multi-factor authentication and conduct regular security audits to prevent data breaches. Diverse and representative data samples will be used to minimize biases, and clustering results will be validated with external benchmarks and qualitative feedback from business stakeholders. Using clear documentation will help prevent misinterpretation of results.

**Continuous Monitoring and Evaluation**: A monitoring system will be established to continuously evaluate the ethical implications and risks associated with the research.

## 8. Description of artefact(s) that will be created (if applicable).

In this project, a few artefacts will be developed to support the research findings:

**Customer Segmentation Model:** A detailed K-means clustering model will be developed to segment B2B customers based on their purchasing behaviors and demographic characteristics and, the evaluation of the accuracy and effectiveness of the segmentation model will be assessed using internal validation metrics such as Silhouette Score and Davies-Bouldin Index.

**Cluster Profiles:** Comprehensive profiles will be created for each customer segment identified by the K-means clustering model. These profiles will include key characteristics, purchasing patterns, and demographic information. The relevance and usefulness of the cluster profiles will be evaluated through qualitative feedback from business stakeholders.

**Visualization Dashboards:** Interactive dashboards will be designed to visualize the clustering results and customer segments. These dashboards will feature charts, graphs, and other visual tools to represent data insights clearly.

**Recommendations:** A set of recommendations will be developed based on the insights gained from the customer segmentation model.

**Technical Documentation**: Detailed documentation of model will be created, including data preprocessing steps, feature selection criteria, clustering process, and validation techniques.

## *9.* **Timeline of proposed activities.**

The Gantt chart illustrates the 12-week timeline for my project, "Evaluating K-means Clustering for Customer Profiling in B2B Ecommerce." The project begins with a literature review during the first two weeks, followed by nine days of research design and planning, including developing the research design, planning data collection, and obtaining necessary permissions. The next three weeks are dedicated to data collection and preprocessing to ensure data uniformity. This is followed by a four-day phase for feature selection and applying the K-means clustering algorithm. Validation and evaluation will occur over six days, where the clustering model will be validated both internally and externally. The subsequent five days will focus on analysing the characteristics and behaviours of each customer segment. In the following seven days, I will develop the segmentation model, create visualization dashboards, draft business recommendations, and prepare technical documentation. The final week is reserved for the review and preparation of the presentation, including finalizing artefacts, preparing slides, the writing and practicing the presentation.

*That’s it from my side! I hope you have enjoyed my presentation and thank you very much for your attention!*

## References

Aarhus University (2022). *IMRAD structure*. [online] studypedia.au.dk. Available at: https://studypedia.au.dk/en/formal-requirements/imrad-structure [Accessed 27 Apr. 2024].

Azevedo, M. de A. and Gartner, I.R. (2020). Concentração e Competição no Mercado de Crédito Doméstico. *Revista de Administração Contemporânea*, [online] 24, pp.380–399. doi:https://doi.org/10.1590/1982-7849rac2020190347.

Borlea, I.-D., Precup, R.-E., Borlea, A.-B. and Iercan, D. (2021). A Unified Form of Fuzzy C-Means and K-Means algorithms and its Partitional Implementation. *Knowledge-Based Systems*, 214, p.106731. doi:https://doi.org/10.1016/j.knosys.2020.106731.

Essayem, W., Bachtiar, F.A. and Priharsari, D. (2022). *Customer Clustering Based on RFM Features Using K-Means Algorithm*. [online] IEEE Xplore. doi:https://doi.org/10.1109/CyberneticsCom55287.2022.9865572.

Golash-Boza, T. (2022). *How to Write a Literature Review: Six Steps to Get You from Start to Finish*. [online] www.wiley.com. Available at: https://www.wiley.com/en-us/network/publishing/research-publishing/writing-and-conducting-research/writing-a-literature-review-six-steps-to-get-you-from-start-to-finish [Accessed 2 Apr. 2024].

Govender, P. and Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, 11(1), pp.40–56. doi:https://doi.org/10.1016/j.apr.2019.09.009.

Huang, S., Kang, Z., Xu, Z. and Liu, Q. (2021). Robust deep k-means: An effective and simple method for data clustering. *Pattern Recognition*, 117, p.107996. doi:https://doi.org/10.1016/j.patcog.2021.107996.

Kotler, P. and Armstrong, G. (2014). *Principles of marketing*. Upper Saddle River, N.J.: Pearson.

Li, Y., Chu, X., Tian, D., Feng, J. and Mu, W. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing*, 113, p.107924. doi:https://doi.org/10.1016/j.asoc.2021.107924.

Lloyd, C. (2017). *LITERATURE REVIEWS FOR SOCIOLOGY SENIOR THESES*. [online] Available at: https://socthesis.fas.harvard.edu/files/socseniorthesis/files/pres-litreview.pdf [Accessed 2 Apr. 2024].

Mensouri, D., Azmani, A. and Azmani, M. (2022). K-Means Customers Clustering by their RFMT and Score Satisfaction Analysis. *International Journal of Advanced Computer Science and Applications*, 13(6). doi:https://doi.org/10.14569/ijacsa.2022.0130658.

Moshkovitz, M., Dasgupta, S., Rashtchian, C. and Frost, N., 2020, November. Explainable k-means and k-medians clustering. In International conference on machine learning (pp. 7055-7065). PMLR. mlr.press

Mussabayev, R., Mladenovic, N., Jarboui, B. and Mussabayev, R. (2022). How to Use K-means for Big Data Clustering? *Pattern Recognition*, p.109269. doi:https://doi.org/10.1016/j.patcog.2022.109269.

Nedyalkova, M., Madurga, S. and Simeonov, V. (2021). Combinatorial K-Means Clustering as a Machine Learning Tool Applied to Diabetes Mellitus Type 2. *International Journal of Environmental Research and Public Health*, 18(4), p.1919. doi:https://doi.org/10.3390/ijerph18041919.

Nouraei, H., Nouraei, H. and Rabkin, S.W. (2022). Comparison of Unsupervised Machine Learning Approaches for Cluster Analysis to Define Subgroups of Heart Failure with Preserved Ejection Fraction with Different Outcomes. *Bioengineering*, 9(4), p.175. doi:https://doi.org/10.3390/bioengineering9040175.

Peppers, D. and Rogers, M. (2010). *Managing Customer Relationships*. John Wiley & Sons.

Practical Data Science (2021). *A quick guide to customer segmentation for B2B e-commerce*. [online] practicaldatascience.co.uk. Available at: https://practicaldatascience.co.uk/data-science/a-quick-guide-to-customer-segmentation-for-b2b-e-commerce#google_vignette [Accessed 27 Apr. 2024].

Sembiring Brahmana, R.W., Mohammed, F.A. and Chairuang, K. (2020). Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, 11(1), p.32. doi:https://doi.org/10.24843/lkjiti.2020.v11.i01.p04.

Srivastava, S., Giri Gundu Hallur and Mukitm, A. (2023). Advanced-data analytics in telecommunications industry: A case study of accenture. *Nucleation and Atmospheric Aerosols*. [online] doi:https://doi.org/10.1063/5.0170732.

Vohra, R., Pahareeya, J., Hussain, A., Ghali, F. and Lui, A. (2020). Using Self Organizing Maps and K Means Clustering Based on RFM Model for Customer Segmentation in the Online Retail Business. *Intelligent Computing Methodologies*, pp.484–497. doi:https://doi.org/10.1007/978-3-030-60796-8_42.

Xiaoyu, Z. (2023). Exploring the Efficacy of Adaptive Learning Technologies in Online Education: A Longitudinal Analysis of Student Engagement and Performance. *International Journal of Science and Engineering Applications*. [online] doi:https://doi.org/10.7753/ijsea1212.1007.

Zare, H. and Emadi, S. (2020). Determination of Customer Satisfaction using Improved K-means algorithm. *Soft Computing*, 24(22), pp.16947–16965. doi:https://doi.org/10.1007/s00500-020-04988-4.