**CHAPTER 3: METHODOLOGY**

This chapter outlines the comprehensive approach undertaken in the data analysis project aimed at optimizing retail operations to gain a competitive advantage and increase sales for Walmart Inc. The methodology encompasses several key stages, including data collection and preparation, exploratory data analysis (EDA), predictive modelling development, implementation strategy formulation, framework development, framework validation, and documentation and knowledge transfer.

**Research Approach and Research Design**: The overall research approach for this project is to utilize secondary data to provide actionable insights that can be implemented by Walmart Inc. to gain a competitive advantage and increase profitability. The rationale for using secondary data is that it provides a cost-effective and time-efficient method for obtaining relevant data, which is essential for achieving the project's objectives. A datset containing secondary data will be sourced from Kaggle. This approach aligns with the research objectives and questions by providing a comprehensive dataset that can be used to conduct a literature review, exploratory data analysis, predictive modelling, and framework development. The secondary data will enable the project to identify patterns, trends, and outliers in Walmart Inc.'s retail operations and sales performance, which can be used to develop predictive models and formulate an implementation strategy. The use of secondary data also allows for the validation of the proposed framework through benchmarking against industry benchmarks, best practices, or competing solutions available in the market (Pederson, 2020). The secondary data will be analyzed using quantitative data analysis and qualitative data analysis, which will provide a comprehensive understanding of the underlying factors impacting Walmart Inc.'s retail operations and sales performance. In conclusion, the use of secondary data is essential for achieving the project's objectives by providing a comprehensive dataset that can be used to conduct a literature review, exploratory data analysis, predictive modelling, and framework development. The secondary data will enable the project to identify patterns, trends, and outliers in Walmart Inc.'s retail operations and sales performance, which can be used to develop predictive models and formulate an implementation strategy. Secondary data also allows for the validation of the proposed framework through benchmarking against industry benchmarks, best practices, or competing solutions available in the market.

**Data Source Description**: The primary data source, WalmartSalesData.csv, obtained from Kaggle, includes variables relevant to Walmart Inc.'s operations, aligning with the project's objectives. It contains data about Sales transactions, customer demographics, store layout, and inventory levels.

**Rationale for Using Secondary Data**: Secondary data is appropriate for this study due to its cost-effectiveness, time efficiency, and access to a large volume of relevant information. As emphasized by (Wickham, 2019), using secondary data is more convenient than gathering primary data, which typically is the most time-consuming and costly component of the research.By leveraging pre-existing data, sourced from datasets such as WalmartSalesData.csv obtained from Kaggle, provide comprehensive insights into various aspects of Walmart Inc.'s operations, including sales transactions, customer demographics, store layout, and inventory levels. Instead of starting from scratch and trying to gather all the data , the information that's already out there can be used. This saves time and effort without worrying about having to find people to take part in the study, whether the collected data is enough, or worrying about people dropping out of the study. Secondary data facilitates a thorough analysis of historical trends and patterns, facilitating the development of data-driven strategies to optimize retail operations and increase sales. Purpose: The use of secondary data is justified by its relevance to the research objectives, enabling a comprehensive analysis of retail operations and sales performance.

**Data Selection Criteria**: The dataset was selected based on criteria such as completeness, relevance, and availability of key variables. However, the limitation of the dataset covering a period of three months is acknowledged.

**Data Analysis Procedures**: Data Analysis was done using Power BI and Excel… Excel was used to obtain the summary of descriptive statistics. They provide insights into the central tendency, variability, and distribution of relevant variables under investigation (Unit Price, Total Sales, Cost of Goods Sold, Gross Income, and Rating). Subsequently, visualizations done on Power BI are used in comparative analysis to identify the impact of independent variables on dependent variables. Comparative/ Benchmarking analysis is another important method used to compare different segments of the dataset or compare Walmart Inc.'s performance with industry benchmarks or competitors. This analysis helps to identify areas of strength and weakness and benchmark performance against industry standards (D'Lima, 2014) . Moreover, qualitative methods such as content analysis may be applied to analyse textual data within the dataset, such as customer feedback or product reviews. Content analysis involves systematically categorizing and interpreting qualitative data to identify recurring themes or patterns, providing additional insights into customer preferences and behaviors. The data analysis procedure aims to interpret the secondary data comprehensively, providing valuable insights into retail operations and sales performance for Walmart.

**Evaluation of Data Quality**: To ensure the integrity of the analysis, the reliability, validity, and limitations of the secondary data are thoroughly assessed. Exploratory Data Analysis (EDA) is conducted to uncover patterns, trends, and outliers in the data (IBM, 2020). Valuable insights into sales and customer behavior are gained through the utilization of various techniques such as visualizations, statistical summaries, and segment analysis. However, EDA plays a crucial role in assessing data quality, identifying anomalies, and generating hypotheses for further analysis (IBM, 2020). It may also have limitations, such as its inability to detect certain complex relationships within the data. Therefore, a comprehensive assessment of data quality involves considering the strengths and weaknesses of EDA alongside other evaluation methods, ensuring a thorough understanding of the dataset's reliability and validity.

**Predictive Modelling Development**: The project proceeds with the development of predictive models using statistical techniques such as Linear regression. Leveraging historical data, these models provide valuable insights for strategic decision-making across different aspects of Walmart Inc’s retail operations. These models aim to forecast consumer demand, optimize inventory levels, and enhance sales forecasting accuracy. Predictive modeling also plays a crucial role in informing pricing strategies for Walmart Inc. By analyzing historical pricing along various market factors as seen in the dataset, predictive models can provide insights into optimal pricing strategies.

**Implementation Strategy Formulation**: A comprehensive strategy is devised for implementing data analytics solutions in Walmart Inc.'s retail operations. Key considerations such as technology selection, organizational change management, collaborations, and resource allocation are addressed to ensure successful implementation and adoption of data analytics practices.

**Framework Development and Validation**: A framework is developed to guide the design and implementation of the project, focusing on areas such as retail operations efficiency, inventory management, and customer experience. The framework's effectiveness and viability are then validated through benchmarking against industry standards and best practices.

**Documentation and Knowledge Transfer**: Detailed documentation and training materials are created to facilitate the adoption and integration of data analytics practices within Walmart Inc.'s personnel. Stakeholders are equipped with the necessary skills and knowledge to leverage data effectively for decision-making, ensuring the successful implementation and sustainability of data-driven initiatives.

**Ethical Considerations**: The confidentiality of the data that has been used in this research has been considered. Each customer has been identified by the invoice ID, gender and customer type. It doesn’t really contain data which can be deemed as sensitive therefore it is safe to say privacy rights have been protected by the person who initially collected the data. Biases have been avoided during the data analysis process to ensure fairness and equity of the individuals represented in the data. This can be seen in the sample data being used for predictive modeling where an equal number of both genders is included to avoid perpetuating stereotypes or discrimination. Transparency has been maintained throughout the project about the practices involved from accessing the WalmartSalesDataset, preprocessing and analysing it as well as on how it was used in predictive model development and the methodologies used.

**Limitations of Secondary Analysis**: The data being used in the project is secondary data obtained from Kaggle, however, conducting secondary analysis mean relying on data collected by others. This lack of control over the data collection process can lead to limitations in the depth of analysis. As seen in the Exploratory Data Analysis, it was identified that the dataset covers a specific period of three months. This limits the ability to analyze long term trends or seasonal variations. There is also a risk of misinterpreting or misunderstanding the context in which the data was collected. Without firsthand knowledge of the data collection process and any limitations of biases that may have been present, it may be possible that incorrect conclusions or faulty assumptions about the data are made about the data during analysis. Another highlighted issue is that Secondary Data Analysis in this project is limited to the variables and information contained in the WalmartSalesdataset.csv. The challenge of not being able to collect primary data to suit the nature of the research and address specific research aims/questions is limiting because some hypothesis cannot be made based on that dataset. When analyzing secondary data, there is a limitation in terms of the scope of available data and certain types of study designs and statistical tests as highlighted by (Wickham, 2019). These studies typically focus on describing, exploring, and finding connections between different things in the data. They rely on observations from past events, so they're often looking backward rather than forward. This means test cause-and-effect relationships wouldn’t be possible unlike in a controlled experiment where changes can be made during data collection to see what happens.

# References

D'Lima, B. K. a. C., 2014. Benchmarking as a measure of competitiveness. *International Journal of Process Management and Benchmarking,* 4(3), p. 342.

IBM, 2020. *IBM.* [Online]   
Available at: https://www.ibm.com/topics/exploratory-data-analysis  
[Accessed 20 March 2024].

Pederson, L. a. V. a. W. a. K. a. R. M., 2020. Use of secondary data analyses in research: Pros and Cons. *Journal of Addiction Medicine and Therapeutics Science,* pp. 058-060.

Wickham, R., 2019. Secondary Analysis Research. *Journal of the Advanced Practitioner in Oncology,* 10(4), pp. 395-400.