**Proposal:**  
**Imputation Techniques for Handling Missing Data in Survey Research**

**Introduction**

Surveys are widely used to collect data from a sample of individuals to understand broader populations. They play a crucial role in various fields, including business, media, government, and academic research, by analyzing public opinion, consumer behavior, and societal trends. In disciplines such as social sciences, healthcare, market research, education, and policy-making, survey research helps organizations make informed decisions. For instance, in the social sciences, surveys provide insights into human behavior and attitudes, while in healthcare, they assess public health trends, patient satisfaction, and medical interventions. Similarly, businesses rely on survey data to analyze consumer preferences and purchasing behaviors, enabling them to refine their products and services. Surveys can be conducted through various methods, including face-to-face interviews, telephone surveys, online questionnaires, and mail-in forms each with its own advantages and challenges in terms of cost, response rates, and data accuracy. Well-designed surveys ensure reliable data collection through structured questions, representative sampling, and appropriate analysis. Khan (2024**)** emphasizes the importance of a systematic approach to survey design to minimize biases and enhance credibility, while (Holtom et al., 2022) highlight the role of tailored survey designs in maximizing response rates and improving data quality.

A significant challenge in survey research is missing data, which can arise from various sources, including non-responses to survey questions, data collection errors, and loss of information due to external factors. This issue can significantly affect the reliability and validity of research findings. Missing data is typically classified into three categories: Missing Completely at Random (MCAR), where the missingness is unrelated to any variable in the study; Missing at Random(MAR), where the missingness is related to an observed variable but not the missing variable itself; and Missing Not at Random (MNAR), where the missingness is directly linked to the missing variable or an unobservable factor (Little, 2021). Failure to address missing data can introduce bias, reduce statistical power, and compromise research validity (Enders, 2022). This issue is particularly prevalent in survey research due to the large number of responses and respondents involved (Mirzaei et al., 2022).

While missing data has been widely studied in fields such as marketing, organizational behavior, and psychometrics (Mirzaei et al., 2022), it has received relatively less attention in mathematical science research. Despite its importance, many studies either ignore missing data or use simplistic methods such as listwise deletion, which can lead to biased results and reduced sample sizes. Addressing this gap, researchers employ imputation techniques to estimate and replace missing values, thereby preserving data integrity and improving the accuracy of statistical analyses.

This study focuses on widely used imputation techniques, including the Average Method, Regression Imputation, K-Nearest Neighbors (K-NN), and Expected Maximization (EM). Rather than providing step-by-step implementation details, the study offers a comprehensive review of these techniques, examining their applications, advantages, and limitations. By understanding the strengths and weaknesses of each method, researchers can make informed choices when handling missing data in survey research.

Imputation is essential for obtaining a complete dataset, but researchers must make several key decisions, including which imputation method to use, how many values to impute for each missing entry, how to select predictor variables, how to handle multivariate nonresponse**,** and how to conduct variance estimation (Bruch, 2023). This study explores the influence of missing data in survey research, emphasizing its impact on statistical analysis and decision-making. By evaluating different imputation methods and their effectiveness, it aims to provide practical recommendations for improving data quality in survey-based studies.

**Statement of Problem**

Missing data in survey research presents significant challenges, including reduced data quality, loss of valuable information, and potential biases in analysis. Missing data occurs when participants do not respond to certain questions or when there are data entry errors, leading to incomplete datasets. This issue is particularly problematic because it can introduce bias, reduce statistical power, and ultimately compromise the integrity of the research findings. Traditional methods of handling missing data, such as listwise deletion or pairwise deletion, may lead to incomplete datasets and distorted conclusions. More advanced techniques, including statistical imputation and machine learning approaches, offer alternative solutions for managing missing values. However, selecting the most appropriate technique requires careful consideration of the nature and extent of missing data. This study aims to examine and evaluate different techniques for handling missing data in survey research, highlighting their advantages and limitations.

**Aim and Objectives**

The aim of this study is to analyze and compare different imputation techniques to determine their efficiency in handling missing data in survey research. The specific objectives are;

* To estimate missing entries using both traditional and advanced techniques for handling missing data based on different categories (MCAR, MNAR, and MAR).
* To evaluate the effectiveness of different imputation techniques (Average Method, Regression Imputation, K-NN, and Expected Maximization) in handling missing data.
* To compare the performance of these techniques based on different categories of the missing data.
* To provide recommendations on the best imputation strategy for different categories of missing data.

**Justification of the Study**

Missing data is a persistent challenge in survey research, affecting the accuracy, reliability, and overall validity of research findings. When responses are incomplete, the resulting data gaps can introduce bias, reduce statistical power, and compromise the generalizability of conclusions. Given that survey research is widely utilized across various fields such as social sciences, healthcare, and market research, it is crucial to adopt appropriate techniques for handling missing data. Without effective imputation methods, researchers may be forced to discard incomplete responses, leading to reduced sample sizes and weakened analytical strength. This study focuses on key imputation techniques, including the Average Method, Regression Imputation, K-Nearest Neighbors (K-NN), and Expectation-Maximization (EM), all of which offer different approaches to estimating missing values while preserving data integrity. By exploring these techniques, this study aims to provide a clear understanding of their effectiveness and guide researchers in choosing the most appropriate method based on their dataset characteristics.

Addressing missing data is not just a technical necessity but also a methodological imperative for ensuring valid and reliable survey results. Although missing data has been extensively examined in fields like statistics and psychometrics, its application in survey-based research still requires more focused attention. Many researchers lack comprehensive guidance on selecting the most suitable imputation techniques for their specific study designs. This research will fill that gap by evaluating and comparing different imputation approaches, offering practical recommendations for researchers dealing with incomplete datasets. Additionally, as data-driven decision-making becomes increasingly central to policymaking, business strategy, and academic research, the ability to handle missing data effectively is more important than ever. By improving the quality of survey data, this study will contribute to the advancement of empirical research, ensuring that missing values do not undermine the credibility and impact of research findings.

**Scope of the Study**

The research will focus on the application of imputation techniques for handling missing data in survey research. It examines the causes, types, and consequences of missing data while emphasizing the importance of effective data handling strategies to ensure research validity and reliability. This study will specifically explore four widely used imputation techniques: the Average Method, Regression Imputation, K-Nearest Neighbors (K-NN), and Expectation-Maximization (EM). In addition, the study will be limited to cases where missing data occurs within individual responses rather than instances of total non-response to a survey. It will analyze missing data classified as Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), as these classifications influence the choice of imputation technique. By comparing their effectiveness, advantages, and limitations, the research aims to provide a comprehensive understanding of their applicability in survey-based research.

**Literature Review**

Researchers have explored various imputation techniques to manage missing values, ensuring that data analysis remains robust and valid. This section presents a detailed literature review on missing data in survey research, the types of missing data, the impact of missing values on research validity, and commonly used imputation techniques.

**Missing Data in Survey Research**

Survey research plays a critical role in data collection across various disciplines, but missing data remains a persistent challenge (Achouch et al., 2022). Missing responses can arise from respondent fatigue, refusal to answer specific questions, data entry errors, or technical issues during data collection (Huang & Wang, 2021). The presence of missing data, if not addressed correctly, can distort research findings, making it imperative to apply effective imputation methods (Afkanpour, 2025).

Several studies emphasize that missing data occurs at varying levels, depending on survey design, respondent behavior, and question sensitivity (Karunarathna, 2024). In large-scale surveys, missing data is often inevitable, requiring researchers to adopt systematic approaches to minimize its impact on data integrity (Adhikari et al., 2022).

**Types of Missing Data**

Missing data is a common issue in research that can arise due to various factors, such as respondent non-cooperation, survey design flaws, or data entry errors. Mirzaei et al. (2022) classify missing data into three main categories:

* **Missing Completely at Random (MCAR):** The probability of missingness is entirely independent of any variables in the dataset. When data is MCAR, the missing values do not introduce systematic bias, and analysis remains valid even without imputation (Zhou et al., 2024).
* **Missing at Random (MAR):** The probability of missingness is related to observed variables but not to the missing variable itself. This type of missing data can be corrected through appropriate imputation techniques (Little, 2021).
* **Missing Not at Random (MNAR):** Missingness depends on the missing values themselves or an unobserved factor. This scenario is the most problematic, as standard imputation techniques may not fully account for the bias introduced (Little, 2021).

Understanding these classifications is crucial for this present study in selecting appropriate imputation methods, as improper handling of missing data can lead to misleading research outcomes.

**Implications of Missing Data on Research Validity**

The presence of missing data can have significant consequences on research outcomes, including:

* **Bias in Statistical Estimates:** Missing data can distort parameter estimates, leading to biased results that do not reflect the true population characteristics (Bowler et al.**,** 2025)
* **Loss of Statistical Power:** Reduced sample size due to missing values weakens the statistical power of tests, making it harder to detect significant relationships (Kong et al., 2022).
* **Decreased Generalizability:** If missing data is non-random (MNAR), findings may not be generalizable to the entire population (Zhou et al., 2024).

Researchers have highlighted that addressing missing data through appropriate imputation techniques enhances research reliability and validity (Woods et al., 2024).

**Techniques for Handling Missing Data**

Several techniques have been proposed in literature for handling missing data, each with its strengths and limitations:

1. **Average (Mean/Median) Imputation**

One of the simplest approaches to handling missing data is replacing missing values with the mean or median of the available data. This method is widely used due to its ease of implementation, but it has significant drawbacks:

1. **Regression Imputation**

Regression imputation involves predicting missing values based on relationships between observed variables (Alwateer et al., 2024). This method is effective when the missing data mechanism is MAR

1. **K-Nearest Neighbors (K-NN) Imputation**

The K-Nearest Neighbors (K-NN) method estimates missing values based on the similarity between observations (Rachdi et al., 2021). This technique is commonly used in machine learning and data analytics

1. **Expectation-Maximization (EM) Algorithm**

The EM algorithm iteratively estimates missing data by maximizing the likelihood function (Aubry et al., 2021). This approach is widely used in handling MAR and MNAR data.

**Evaluating Imputation Techniques in Survey Research**

Various studies have compared imputation techniques to determine their effectiveness in handling missing survey data. Chiu et al. (2022) suggests that no single method is superior in all scenarios, emphasizing the importance of selecting imputation techniques based on the nature of missingness.

A comparative study by (Alwateer et al., 2024) found that K-NN and EM imputation performed better than mean imputation in preserving data structure. Similarly, (Gomer & Yuan, 2023). demonstrated that regression imputation works well when MAR assumptions hold, but may introduce bias in MNAR data.

Recent research supports hybrid approaches, combining multiple imputation techniques to improve data accuracy (Nadimi-Shahraki et al., 2021). For example, combining regression imputation with EM enhances estimation accuracy while maintaining computational efficiency (Adhikari et al., 2022).

**Methodology**

This study will adopt a systematic approach to examining imputation techniques for handling missing data in survey research. The methodology will involve a combination of literature review, data simulation, and comparative analysis of different imputation methods. The study will be conducted in the following phases:

**1. Research Design**

The research will be a quantitative, comparative study that evaluates the effectiveness of different imputation techniques in handling missing data. The study will focus on four widely used methods:

* **Average Method**
* **Regression Imputation**
* **K-Nearest Neighbors (K-NN)**
* **Expectation-Maximization (EM)**

**2. Data Collection**

The study will utilize secondary survey datasets obtained from publicly available research repositories. In cases where no suitable dataset is found, synthetic data will be generated to mimic real-world survey responses. These datasets will contain various types of missing data, including:

* **Missing Completely at Random (MCAR)**
* **Missing at Random (MAR)**
* **Missing Not at Random (MNAR)**

The proportion of missing data will be systematically varied to examine how different imputation techniques perform under different conditions.

**3. Data Processing and Handling of Missing Values**

The dataset will undergo preprocessing to ensure consistency and standardization before applying imputation techniques. The study will compare the impact of different imputation methods based on key evaluation criteria, including accuracy of estimated missing values, preservation of statistical relationships in the dataset and impact on overall data distribution. Statistical package such as **R or SPSS** will be considered to aid analysis.

**4. Data Analysis and Evaluation**

To evaluate the performance of each imputation technique, the study will employ statistical metrics such as:

* **Mean Absolute Error (MAE)** – to measure the accuracy of imputed values
* **Root Mean Square Error (RMSE)** – to assess deviation from true values
* **Bias Analysis** – to determine whether imputation introduces systematic errors

Comparative analysis will be conducted to determine which method produces the most reliable and unbiased estimates for different types and proportions of missing data

**Expected Outcome**

One of the key anticipated outcomes is identifying which imputation method best preserves the accuracy and statistical integrity of survey datasets. It is expected that more advanced techniques, such asExpectation-Maximization (EM) and K-NN, will perform better in cases where data is missing at random (MAR) or missing not at random (MNAR), while simpler methods like the Average Method may introduce biases or distort statistical relationships. The expected findings will not only improve the reliability of survey-based research but also assist policymakers, businesses, and academics in ensuring that incomplete data does not compromise the validity of their conclusions.

**References**

Achouch, M., Dimitrova, M., Ziane, K., Sattarpanah Karganroudi, S., Dhouib, R., Ibrahim, H., & Adda, M. (2022). On predictive maintenance in industry 4.0: Overview, models, and challenges. *Applied Sciences*, *12*(16), 8081.

Adhikari, D., Jiang, W., Zhan, J., He, Z., Rawat, D. B., Aickelin, U., & Khorshidi, H. A. (2022). A comprehensive survey on imputation of missing data in internet of things. *ACM Computing Surveys*, *55*(7), 1-38.

Afkanpour, M., Dehkordy, D. T., Momeni, M., & Tabesh, H. (2025). Conceptual framework as a guide to choose appropriate imputation method for missing values in a clinical structured dataset. *BMC Medical Research Methodology*, *25*, 43.

Alwateer, M., Atlam, E. S., Abd El-Raouf, M. M., Ghoneim, O. A., & Gad, I. (2024). Missing data imputation: A comprehensive review. *Journal of Computer and Communications*, *12*(11), 53-75.

Aubry, A., De Maio, A., Marano, S., & Rosamilia, M. (2021). Structured covariance matrix estimation with missing-(complex) data for radar applications via expectation-maximization. *IEEE Transactions on Signal Processing*, *69*, 5920-5934.

Bowler, D. E., Boyd, R. J., Callaghan, C. T., Robinson, R. A., Isaac, N. J., & Pocock, M. J. (2025). Treating gaps and biases in biodiversity data as a missing data problem. *Biological Reviews*, *100*(1), 50-67.

Chiu, P. C., Selamat, A., Krejcar, O., Kuok, K. K., Bujang, S. D. A., & Fujita, H. (2022). Missing value imputation designs and methods of nature-inspired metaheuristic techniques: A systematic review. *IEEE Access*, *10*, 61544-61566.

Christian Bruch (2023). Imputation of missing values in survey data. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines).

Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.

Gomer, B., & Yuan, K. H. (2023). A realistic evaluation of methods for handling missing data when there is a mixture of MCAR, MAR, and MNAR mechanisms in the same dataset. *Multivariate Behavioral Research*, *58*(5), 988-1013.

Holtom, B., Baruch, Y., Aguinis, H., & A Ballinger, G. (2022). Survey response rates: Trends and a validity assessment framework. *Human relations*, *75*(8), 1560-1584.

Huang, J. L., & Wang, Z. (2021). Careless responding and insufficient effort responding. In *Oxford Research Encyclopedia of Business and Management*.

Karunarathna, I., Gunasena, P., Hapuarachchi, T., & Gunathilake, S. (2024). The crucial role of data collection in research: Techniques, challenges, and best practices. *Uva Clinical Research*, 1-24.

Khan, M. M. (2024). Optimizing web surveys in research: methodological considerations and validity aspects. *International Journal of Research and Scientific Innovation*, *11*(4), 75-105.

Kong, W., Hui, H. W. H., Peng, H., & Goh, W. W. B. (2022). Dealing with missing values in proteomics data. *Proteomics*, *22*(23-24), 2200092.

Little, R. J. (2021). Missing data assumptions. *Annual Review of Statistics and Its Application*, *8*(1), 89-107.

Mirzaei, A., Carter, S. R., Patanwala, A. E., & Schneider, C. R. (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *18*(2), 2308-2316.

Nadimi-Shahraki, M. H., Mohammadi, S., Zamani, H., Gandomi, M., & Gandomi, A. H. (2021). A hybrid imputation method for multi-pattern missing data: A case study on type II diabetes diagnosis. *Electronics*, *10*(24), 3167.

Rachdi, M., Laksaci, A., Kaid, Z., Benchiha, A., & Al‐Awadhi, F. A. (2021). k‐Nearest neighbors local linear regression for functional and missing data at random. *Statistica Neerlandica*, *75*(1), 42-65.

Rea, L. M., & Parker, R. A. (2014). *Designing and conducting survey research: A comprehensive guide*. John Wiley & Sons.

Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., ... & Elsherif, M. M. (2024). Best practices for addressing missing data through multiple imputation. *Infant and Child Development*, *33*(1), e2407.

Zhou, Y., Aryal, S., & Bouadjenek, M. R. (2024). Review for Handling Missing Data with special missing mechanism. *arXiv preprint arXiv:2404.04905*.