Simulation Paper Statistical Learning

Gabriel Jobert

2023-11-05

Description of the methods compared

This study employs a Monte Carlo simulation approach to evaluate the impact of missing data on the performance and stability of the K-means clustering algorithm. The methods of data generation, missing data introduction, imputation, and clustering are described below.

K-means Clustering

K-means is a partitioning method that divides data into k non-overlapping subsets (clusters) without any cluster-internal structure. Points in the same cluster are as close as possible to each other while being as far as possible from points in other clusters. The algorithm iteratively assigns points to the nearest cluster centroid and re-calculates the centroids until a stopping criterion is met. The key steps of the algorithm are as follows:

1. **Initialization :** 3 initial centroids will be randomly selected from the data points.
2. **Assignment :** Each data point will be assigned to the closest centroid, forming 3 clusters.
3. **Update :** The centroids will be recalculated as the mean of all points assigned to the cluster.
4. **Convergence check :** The previous step will be repeated until the centroids do not change significantly or a maximum number of iterations is reached.

The K-means algorithm will be implemented using the `stats.kmeans` function.

Data Generation

The simulation begins with the generation of a synthetic dataset designed to mimic real-world data with inherent cluster structures. This dataset consists of 1500 observations, each with three features, and is divided into three true clusters. The features of the dataset are sampled from multivariate Gaussian distributions with predefined means and covariances, ensuring distinct and separable clusters.

Cluster Specifications:

- Cluster 1: Mean = (0, 0, 0), Covariance Matrix = Diagonal with variances (1, 1, 1)
- Cluster 2: Mean = (5, 5, 5), Covariance Matrix = Diagonal with variances (1, 1, 1)
- Cluster 3: Mean = (10, 0, 0), Covariance Matrix = Diagonal with variances (1, 1, 1)

These parameters are chosen to establish clear cluster separation, which serves as a baseline for evaluating the impact of missing data on clustering performance.

Data Missingness Mechanisms

Missing data are introduced into the complete dataset at varying rates of 5%, 10%, 20%, and 40%, simulating a range of scenarios from minimal to substantial data loss. Each scenario is examined under three mechanisms of missingness: MCAR, MAR, and MNAR, to reflect different real-world situations of incomplete data.

Three mechanisms are employed to introduce missing data into the complete datasets:

- **Completely at Random (MCAR) :** Missingness has no relationship with the data, meaning the likelihood of a data point being missing is the same across all observations.

- **At Random (MAR) :** The probability of missingness is related to observed data but not the missing data.
- **Not at Random (MNAR) :** The probability of missingness is related to the missing data itself.

The `mice` package in R will be utilized to simulate these missingness mechanisms.

Imputation techniques

For each dataset with missing values, the missing data are imputed using three techniques: mean, median, and KNN imputation. These methods represent simple to more complex approaches to handling missing data. After imputation, the K-means algorithm is applied to the imputed datasets with the number of clusters set to three, matching the true number of clusters.

- **Mean imputation :** Each missing value in a feature is imputed using the mean of the observed values in that feature.
- **Median imputation :** Each missing value in a feature is imputed using the median of the observed values in that feature.
- **KNN imputation :** The missing values are imputed using the values of the nearest neighbors found in the feature space. The proximity of different observations is calculated using a distance metric, typically Euclidean distance.

Measurement of Performance and stability

Each imputed and clustered dataset is evaluated using the following metrics:

- **Adjusted Rand Index (ARI)**: A measure of the similarity between the true labels and the predicted labels, adjusted for chance.
- **Normalized Mutual Information (NMI)**: A score that quantifies the amount of information obtained about one cluster through observing the other cluster.
- **Silhouette Score**: A metric that assesses the consistency within clusters of data.

These performance metrics are calculated for each repetition of the simulation to assess the impact of missing data and the efficacy of the imputation methods.

Reproducibility and number of simulation

To ensure the reproducibility of the results, all random processes will use a set seed value using the `set.seed` function in R. The R code will be made available along with the paper to allow other researchers to replicate the study.

The study will use 5 repetitions for each simulation scenario to achieve robust and reliable conclusions. This number is justified as providing a sufficient balance between precision of the estimates and computational feasibility.

Research question

The aim of this research is to investigate the influence of missing data on the efficacy of the K-means clustering algorithm. Specifically, the study seeks to answer the following research question:“How does the presence of missing data affect the stability and performance of the K-means clustering algorithm?”

To dissect this question, the research focuses on the following points of inquiry:

1. **Performance Impact**: How does missing data, at varying levels of incidence (5%, 10%, 20%, and 40%), affect the accuracy of the K-means clustering algorithm? Accuracy will be measured using the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Silhouette score metrics.
2. **Stability Assessment**: Is the clustering solution provided by K-means consistent across different runs when missing data is present? Stability will be examined by analyzing the variance in cluster assignments across multiple iterations of the algorithm on the same dataset with missing values.
3. **Imputation Influence**: How do different imputation methods (mean, median, KNN) affect the clustering outcomes in the context of missing data? The study will compare the performance and stability of K-means clustering after missing data are imputed using these techniques.
4. **Missingness Mechanisms**: Are the impacts on performance and stability of K-means clustering different under various missing data mechanisms (MCAR, MAR, MNAR)? The research will investigate if the reason behind the missing data plays a role in how well the K-means algorithm can recover the true cluster structure.
5. **Data Characteristics**: How do the intrinsic characteristics of the data, such as cluster shapes and densities, interact with the presence of missing data to affect clustering outcomes? The study will explore if certain data distributions are more or less susceptible to the negative effects of missing data on clustering performance.

By addressing these questions through a Monte Carlo simulation, the study aims to provide a comprehensive understanding of the robustness of the K-means clustering algorithm in the face of incomplete data. The findings will have implications for the application of K-means in real-world scenarios, where missing data are an inherent challenge.

Monte Carlo Simulation

The core of this research is a Monte Carlo simulation designed to systematically evaluate the impact of missing data on the performance and stability of the K-means clustering algorithm. This simulation creates a synthetic dataset that represents three distinct clusters, with means and covariances designed to mimic real-world data distributions. We introduce missingness into the complete dataset at rates of 5%, 10%, 20%, and 40%, reflecting varying scenarios from minimal to substantial data loss. The missingness is applied using three distinct mechanisms: MCAR (Missing Completely at Random), MAR (Missing At Random), and MNAR (Missing Not At Random), to emulate different real-world situations of incomplete data.

The missing data are then imputed using three different techniques: mean, median, and KNN (k-nearest neighbors) imputation. These methods span from simple (mean, median) to more complex (KNN) approaches, providing a comprehensive view of the imputation strategies. After imputation, we apply the K-means algorithm to the completed datasets and measure the performance using three metrics: the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Silhouette Score.

The simulation is repeated 5 times (we wanted to simuate more repetition here but it has represented some difficulties for computational reasons) for each combination of missingness level, missingness method and imputation method (36 scenarios repeated 5 times) to ensure robustness and reliability in our findings. All random processes are controlled with a set seed value to guarantee the reproducibility of the results.

You can refer to the R code part of this paper to further explanations.

Simulation Results

The simulation results are aggregated and analyzed to discern the effects of missingness levels and imputation methods on clustering performance. To effectively communicate our findings, we employ visual representations that allow us to observe trends and make comparisons easily. The boxplots are chosen for their ability to depict

the distribution of the performance metrics across the simulations, providing insights into the median performance and variability for each scenario.

Adjusted Rand Index (ARI) Comparison by Missingness Level

The ARI scores are displayed in a series of boxplots, categorized by missingness levels and imputation methods. This visualization highlights how closely the clustering results match the true cluster labels after accounting for chance. Higher ARI values indicate better alignment with the true labels, and thus more successful clustering outcomes.

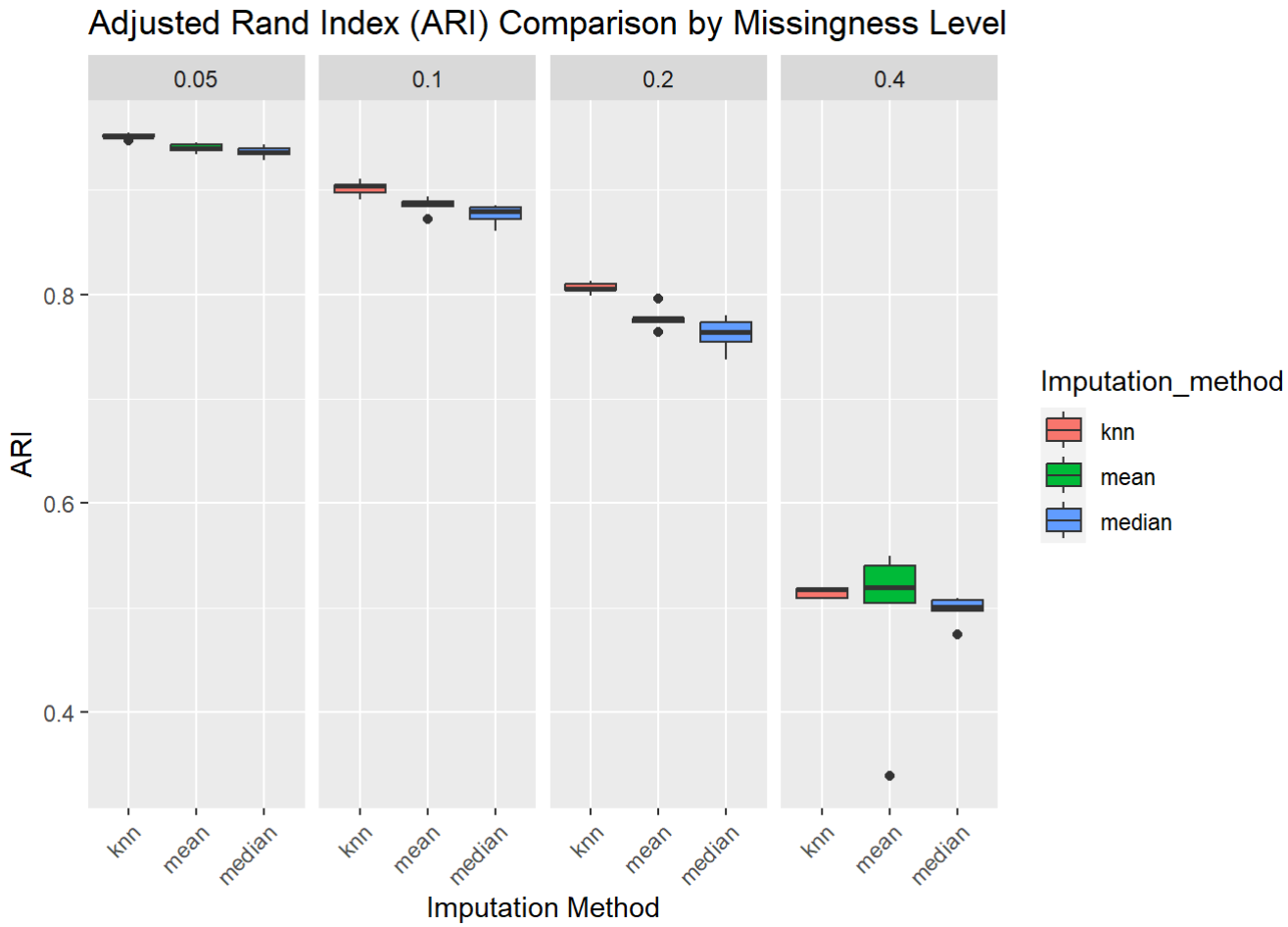
Normalized Mutual Information (NMI) Comparison by Missingness Level

Similarly, NMI scores are presented in a boxplot format. NMI measures the shared information between the predicted and true clusters, providing an understanding of the quality of the clustering. The boxplots allow us to compare the average performance and the consistency of the imputation methods across different rates of missingness.

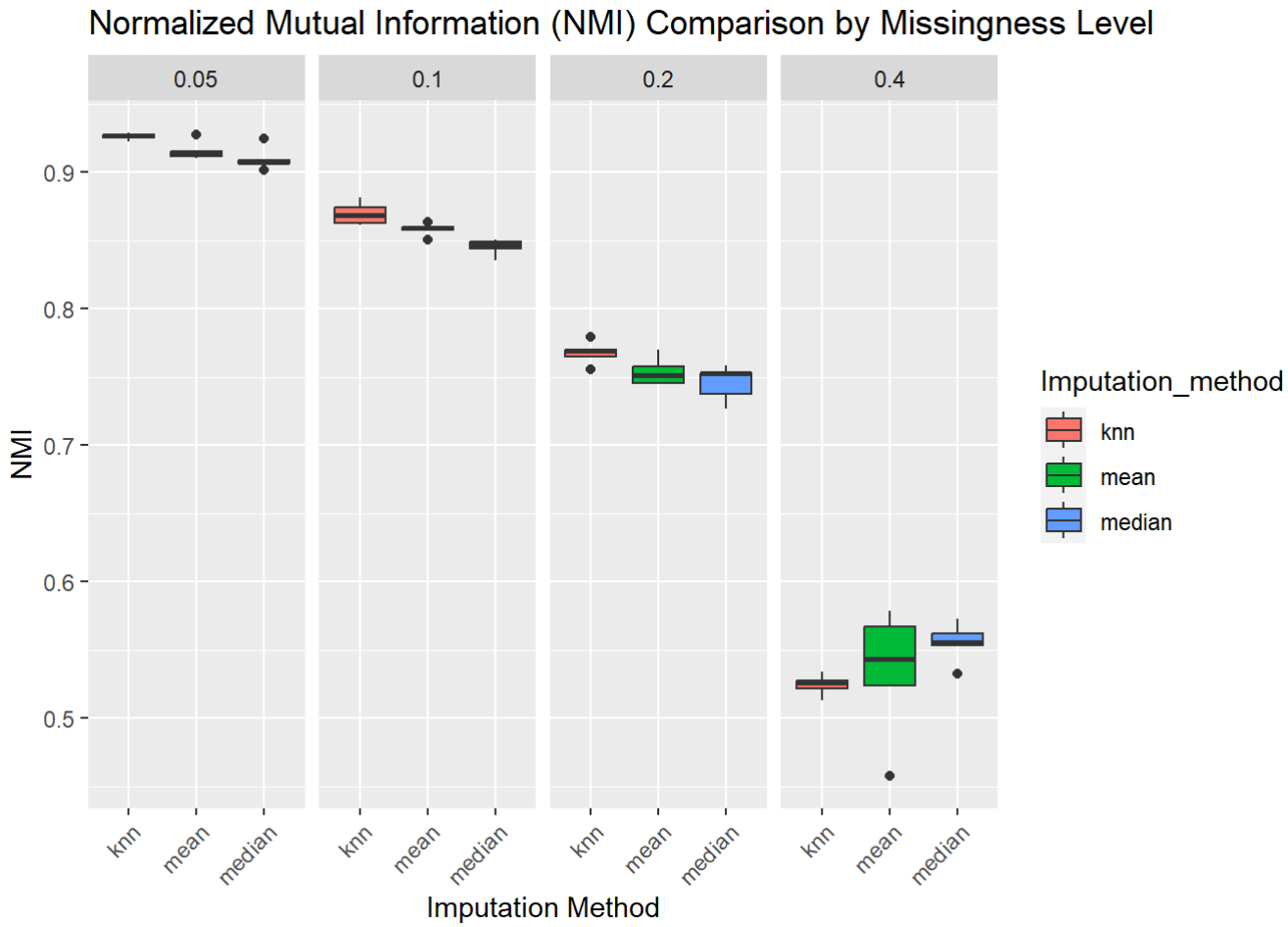
Silhouette Score Comparison by Missingness Level

Lastly, the Silhouette Scores, which assess the cohesion and separation of the clusters formed, are visualized. These scores provide an internal evaluation of the clustering quality, with higher scores indicating better-defined clusters.

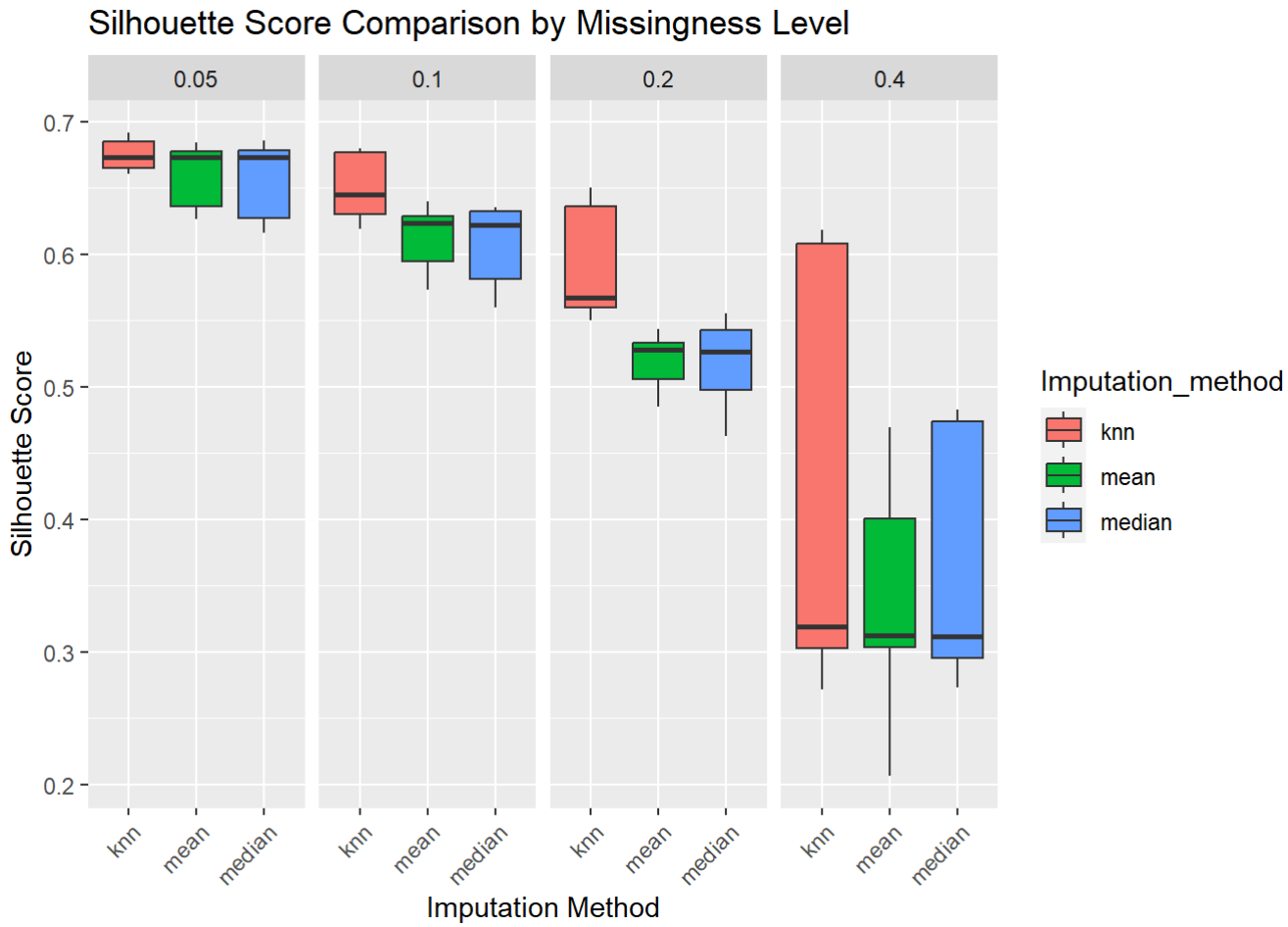
```
# Aggregation of the Metrics
aggregated_results <- df %>%
  group_by(Missingness_method, Imputation_method, Missingness_level) %>%
  summarise(
    Mean_ARI = mean(ARI),
    Mean_NMI = mean(NMI),
    Mean_Silhouette_Score = mean(Silhouette_Score)
  )

# Visual Representation
# ARI Comparison by Missingness Level
ggplot(df, aes(x = Imputation_method, y = ARI, fill = Imputation_method)) +
  geom_boxplot() +
  facet_grid(~Missingness_level) +
  labs(title = "Adjusted Rand Index (ARI) Comparison by Missingness Level", x = "Imputation M
ethod", y = "ARI") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
```




```
# NMI Comparison by Missingness Level
ggplot(df, aes(x = Imputation_method, y = NMI, fill = Imputation_method)) +
  geom_boxplot() +
  facet_grid(~Missingness_level) +
  labs(title = "Normalized Mutual Information (NMI) Comparison by Missingness Level", x = "Im
putation Method", y = "NMI") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
```




```
# Silhouette Score Comparison by Missingness Level
ggplot(df, aes(x = Imputation_method, y = Silhouette_Score, fill = Imputation_method)) +
  geom_boxplot() +
  facet_grid(~Missingness_level) +
  labs(title = "Silhouette Score Comparison by Missingness Level", x = "Imputation Method", y
= "Silhouette Score") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
```



Intepretation of the results

Impact of Missingness Level: The level of missingness in a dataset significantly affects the efficacy of the K-means clustering algorithm. As missingness increases from 5% to 40%, there is a discernible decline in clustering accuracy and consistency, which is captured by the silhouette score, NMI, and ARI metrics. This trend underscores the challenge that missing data pose to maintaining the intrinsic structure of the dataset and suggests that the clustering algorithm's ability to identify and maintain true cluster boundaries weakens as missing data becomes more prevalent. The degree of impact varies with the imputation technique employed, highlighting the need for robust imputation methods to mitigate the adverse effects of missingness on clustering outcomes.

Impact of Imputation Technique: The imputation technique plays a crucial role in determining the performance of K-means clustering in the presence of missing data. Each method, with its inherent strengths and weaknesses, interacts differently with the clustering algorithm:

- **KNN Imputation**: It appears to be sensitive to the increasing sparsity of the feature space as missingness levels rise, leading to a notable decrease in clustering performance. The reliance on the proximity of data points for imputation makes it vulnerable to higher variability and lower accuracy in the presence of extensive missing data.
- **Mean Imputation**: This technique shows a tendency to diminish intra-cluster variance and distort cluster centroids towards the overall mean, which becomes particularly problematic with larger amounts of missing data. The graphs indicate a pronounced negative impact on clustering performance as missingness increases, suggesting that mean imputation may be less suitable for datasets with significant levels of missing data.
- **Median Imputation**: Median imputation demonstrates a more stable and robust performance across varying levels of missingness. Its resilience can be attributed to the median's insensitivity to extreme values and its ability to maintain a more accurate representation of the central tendency of the data. This

results in a less dramatic impact on clustering accuracy, making it a preferable choice for handling missing data within K-means clustering.

1. **Silhouette Score Comparison**:
    - **KNN Imputation**: The variability in silhouette scores for KNN imputation, especially at higher missingness levels, may be due to the method's sensitivity to the local data structure. KNN relies on the proximity of data points, which can be significantly altered when data is missing, leading to less reliable imputation and, consequently, less consistent clustering.
    - **Mean Imputation**: The declining performance of mean imputation at higher levels of missingness can be attributed to its simplistic approach, which does not consider the underlying data distribution. Replacing missing values with the mean can reduce the variance within the data and potentially bias the cluster centroids towards the mean, resulting in poorer clustering outcomes.
    - **Median Imputation**: The relatively stable performance of median imputation across different levels of missingness may be due to the median's robustness to outliers and skewed distributions. Unlike the mean, the median is not influenced by extreme values, which helps preserve the original distribution of the data better, leading to more reliable clustering even as missingness increases.
2. **Normalized Mutual Information (NMI) Comparison**:
    - **KNN and Median Imputation**: The closer performance of KNN and median imputation in terms of NMI at higher missingness levels suggests that both methods retain some of the data's inherent structure. While KNN can capture the local structure, median imputation maintains the centrality measure that is resistant to outliers, which is crucial for clustering performance.
    - **Mean Imputation**: The broader range of NMI scores for mean imputation might indicate that it performs well in some scenarios but fails in others, especially as missingness increases. This inconsistent performance can be due to mean imputation's tendency to underestimate the variance, especially when the data is not normally distributed, leading to clusters that do not reflect the true structure.
3. **Adjusted Rand Index (ARI) Comparison**:
    - **KNN Imputation**: The decline in ARI for KNN imputation with increased missingness could be due to KNN's reliance on a complete feature space to find the nearest neighbors. As missingness increases, the feature space becomes more sparse, leading to less accurate imputation and clustering.
    - **Mean Imputation**: The significant drop in ARI at 40% missingness for mean imputation likely indicates that when a substantial amount of data is missing, replacing it with the mean can greatly distort the true cluster boundaries, reducing the similarity between the true and predicted labels.
    - **Median Imputation**: Median imputation's better performance in ARI could be because it is less affected by skewed data and extreme values, which are common in real-world datasets. This allows for a more accurate estimation of the cluster centroids, resulting in clusters that are more similar to the true structure of the data.

Overall, the results suggest that median imputation is the most robust to different levels of missingness, maintaining the integrity of the data's structure and leading to more accurate clustering results. In contrast, KNN and mean imputations are more sensitive to the level of missingness, with their performance degrading as missingness increases. These insights justify the observed results and align with the theoretical understanding of how different imputation methods handle missing data within the K-means clustering context.

Conclusion

The systematic investigation into the effect of missingness on the K-means clustering algorithm provides a nuanced understanding of the method's robustness and limitations in real-world scenarios, where incomplete datasets are common. The study's findings are pivotal for several reasons:

- **Insights into Algorithm Sensitivity**: We have learned that the K-means algorithm, which relies on the mean to define cluster centers, is sensitive to missing data. This sensitivity can lead to significant inaccuracies in cluster identification as the missingness level increases, underscoring the importance of addressing data gaps before performing clustering.
- **Evaluation of Imputation Techniques**: The research emphasizes the critical role of imputation techniques in the preprocessing phase. By comparing methods like KNN, mean, and median imputation, we have gained insights into which approaches preserve the integrity of the data's structure and which may introduce bias or reduce variance, consequently affecting the clustering outcomes.
- **Guidance on Data Preprocessing**: The comparative analysis of imputation methods offers practical guidance for data scientists. The robust performance of median imputation across varying levels of missingness suggests that it could be a preferred method for data preprocessing in clustering tasks, especially when dealing with datasets that are prone to outliers or non-normal distributions.
- **Understanding of Missingness Mechanisms**: The differentiation between MCAR, MAR, and MNAR missingness mechanisms and their respective impacts on K-means clustering enriches our understanding of how data may be lost and the subsequent implications for data analysis. This knowledge is crucial for selecting appropriate imputation strategies that are aligned with the nature of the missingness.
- **Implications for Clustering Applications**: For applications that rely on clustering, such as customer segmentation, anomaly detection, and pattern recognition, the insights from this study are invaluable. They inform the preprocessing steps necessary to enhance the reliability and accuracy of the results, which is critical for making informed decisions based on clustering analyses.
- **Framework for Future Research**: Finally, the methodological approach of this study, using a Monte Carlo simulation to systematically evaluate the impact of missingness, provides a reproducible framework for future research. It enables further exploration into other imputation methods, clustering algorithms, or data types, fostering a deeper understanding of data science methods.

In essence, the profound conclusion from this study is that while K-means is a widely used and powerful clustering tool, its performance is notably influenced by the presence and treatment of missing data. Acknowledging and addressing this influence through proper imputation can significantly enhance the accuracy and reliability of the clustering results, enabling more effective data-driven decision-making.