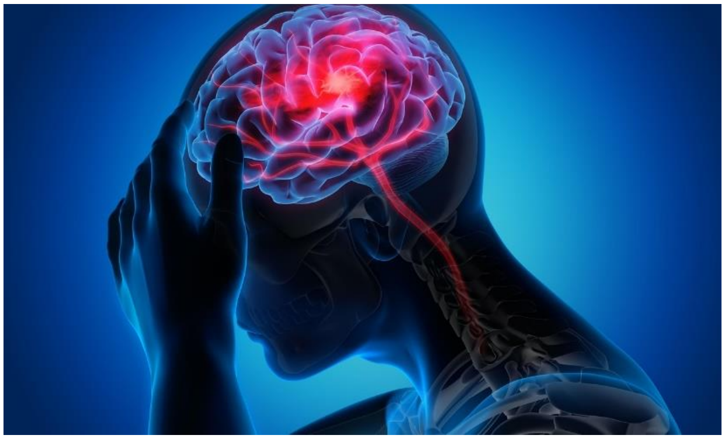


MIGRAINE SYMPTOM CLASSIFICATION

Exploratory Data Analysis

Supul Wicramasinghe – s16177

Thenuka Yatawara – s16383

Vayani Kavindya - s16322

Sanduni Fonseka – s16026

ABSTRACT

This project presents an exploratory data analysis (EDA) of a migraine symptom classification dataset, aimed at uncovering meaningful patterns and relationships among demographic, clinical, and symptom-based variables. The dataset includes a wide range of features such as migraine type, symptom presence, pain intensity, frequency, duration, and age.

Through a combination of univariate, bivariate, and multivariate analysis techniques—along with dimensionality reduction and clustering methods like FAMD—we explore the structure and variability within the data. Our analysis reveals strong associations between certain symptoms and migraine subtypes, highlights imbalances in symptom distribution, and uncovers potential symptom-based groupings of patients. These insights provide a foundation for improved understanding of migraine characteristics and support future clinical decision-making or predictive modeling approaches.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Migraines are a complex neurological disorder characterized by recurrent headaches that typically cause severe throbbing or pulsing pain, often on one side of the head. These headaches are frequently accompanied by symptoms such as nausea, vomiting, and heightened sensitivity to light and sound. Some migraine sufferers experience an aura—a set of sensory disturbances including visual flashes, blind spots, tingling sensations, or speech difficulties—that can precede or accompany the headache. Migraines can last from several hours to days and can significantly disrupt daily life and activities.

Understanding the variations in migraine symptoms and their patterns is clinically important because it aids in accurate diagnosis and effective treatment planning. Different types of migraines manifest with distinct symptom profiles, and a detailed examination of these symptoms can help healthcare professionals tailor management strategies to individual patients.

The purpose of this project is to perform an exploratory data analysis on a migraine classification dataset. The goal is to uncover meaningful patterns, trends, and relationships in the symptom data without applying predictive modeling. By analyzing the distribution and association of symptoms across different migraine types, this study aims to provide insights that can support better clinical understanding and decision-making.

The report is organized as follows: a description of the question, a description of the dataset, the main results from descriptive and visual analyses, followed by suggestions for future advanced analysis. The appendix contains the technical details and code used for this study.

DESCRIPTION OF QUESTION

Migraine is a common and debilitating neurological disorder that affects millions of individuals worldwide. Despite its prevalence, the condition is often underdiagnosed or misdiagnosed due to the wide variety of symptoms that patients experience. Migraines can differ significantly in intensity, location, associated sensory disturbances, and presence or absence of aura. This variation makes it difficult for clinicians to quickly and accurately classify the type of migraine a patient may have, especially when symptoms overlap between types.

In clinical settings, understanding how certain symptoms co-occur and how they vary across different types of migraines can aid in faster diagnosis and better-targeted treatments. However, in most cases, patients are diagnosed based on subjective symptom reporting, and there is little exploration of broader patterns in symptom combinations.To address this challenge, our project aims to explore the following key question:

How do clinical symptoms and pain characteristics vary across different types of migraines, and can these variations reveal distinctive symptom profiles for each type?

By conducting a detailed exploratory data analysis of a migraine symptom dataset, we aim to uncover meaningful trends and patterns in symptom distribution. We focus on analyzing the relationships between migraine types and symptoms such as nausea, vomiting, pain intensity, pain location and nature, visual and sensory disturbances, and other aura-related features. Identifying these patterns may provide insights that could contribute to improving early diagnosis and clinical decision-making in migraine care.

DESCRIPTION OF THE DATA SET

The dataset consists of detailed clinical and demographic information on patients diagnosed with different types of migraines. It was obtained from Kaggle and originally contained 400 observations and 24 variables. After identifying and removing 6 duplicate records, the final dataset includes 394 unique patient records. Initially, all variables were of type int64, but during preprocessing, appropriate variables were converted to categorical types to better reflect their clinical meanings (e.g., symptom presence or pain characteristics).

Table 1:Description of Variables

Variable	Description	Original Type	Final Type
Age	Age of the participant	int64	int64
Duration	Duration of the migraine attack (in hours)	int64	int64
Frequency	Frequency of attack over a specific period	int64	int64
Location	Location of the pain: 1.Unilateral, 2. Bilateral, 3.Frontal	int64	Category
Character	Nature of pain: 1 - Throbbing, 2 - Pressing, 3 - Sharp	int64	Category
Intensity	Pain severity 1. No Pain, 2. Mild, 3. Moderate, 4. Serve	int64	Category
Nausea	Presence of nausea during the episode	int64	Category
Vomit	Presence of Vomiting	int64	Category
Phonophobia	Sensitivity to sound	int64	Category
Photophobia	Sensitivity to light	int64	Category
Visual	Visual disturbances such as aura or light flashes	int64	Category
Sensory	Sensory issues such as numbness or tingling	int64	Category
Motor	Impaired motor function	int64	Category
Language	Language disturbances	int64	Category
Vertigo	Presence of dizziness or vertigo	int64	Category
Tinnitus	Ringing in the ears	int64	Category
Hypoacusis	Decreased hearing ability	int64	Category
Diplopia	Double vision	int64	Category
Defect	Visual field defect or scotoma	int64	Category
Ataxia	Balance or coordination problems	int64	Category
Conscience	Episodes of reduced consciousness	int64	Category
Paresthesia	Abnormal skin sensations such as tingling	int64	Category
DPF	Dysfunctional physical factor	int64	Category
Type	Type of migraine diagnosed, e.g., "Typical aura with migraine". *(Target variable)*	Object	Category

MAIN RESULTS OF THE DESCRIPTIVE ANALYSIS

To better understand the structure and characteristics of the dataset, we performed a series of descriptive and visual analyses on all variables.

Distribution of response variable.

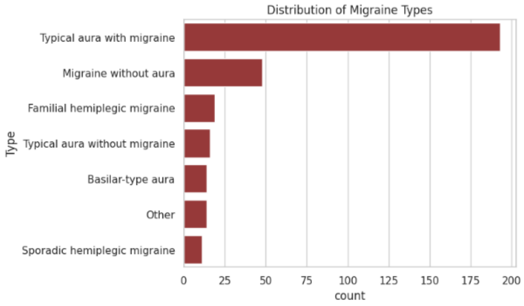


Figure 1:Distribution of Migraine types

Our response variable of interest is the **type of migraine**, which is classified into seven distinct categories. Figure 1 shows that the most common type in our dataset is *Typical aura with migraine*, accounting for more than half of all cases. This is followed by *Migraine without aura*, while other types, such as *Familial hemiplegic migraine*, *Basilar-type aura*, and *Sporadic hemiplegic migraine* are much less frequent. This imbalance in class distribution suggests that most of the findings from our analysis may reflect the characteristics of the most common types.

Numerical Variables – Age, Duration & Frequency

We examined the distribution of the three numerical variables using boxplots and histograms (Figure 2 and Figure 3).

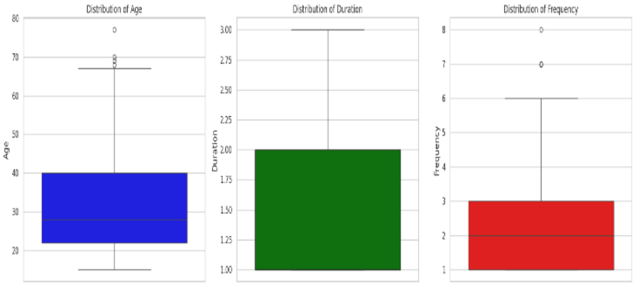


Figure 3: Distribution Boxplots of Numerical variables

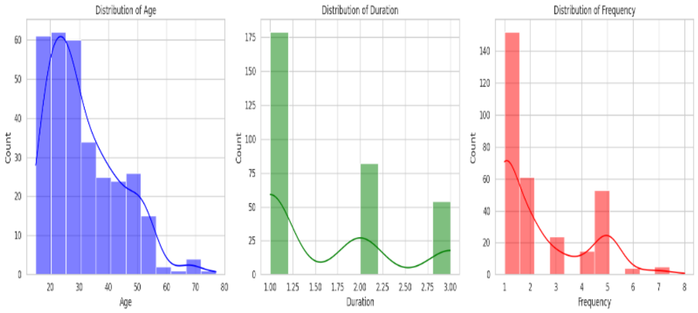


Figure 2: Distribution Histograms of Numerical Variables

Age
The age of patients ranges from 15 to 77 years, with a median around 28 years. The distribution is **right-skewed**, indicating that most migraine sufferers in the dataset are young adults (ages 20–30), while a few older patients contribute to long upper tails and visible outliers in the boxplot.

Duration
This variable represents the duration of migraine attacks, categorized into 1, 2, or 3 (assumed to represent hours or categories). The distribution is **left-skewed**, with **shorter attacks (duration = 1)** being the most common. Outliers are not evident, but the variable shows a strong peak at the lowest category.

Frequency
The frequency of attacks shows a highly **right-skewed** distribution. Most patients experience **1 or 2 attacks** in a given period, with very few reporting frequent episodes (values of 6, 7, or 8). A few outliers are visible in the boxplot. This suggests that while occasional migraines are typical, a subset of individuals suffers from high-frequency migraines.

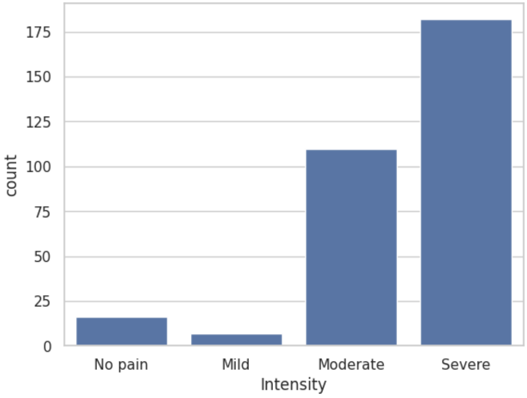


Figure 4: Distribution of migraine pain Intensity levels

Intensity is an ordinal categorical variable representing migraine pain severity with levels: *No pain*, *Mild*, *Moderate*, and *Severe*. As shown in Figure X, the majority of cases report **Severe** pain, while *No pain* is entirely absent. This reflects a **heavily right-skewed and imbalanced distribution**, with the **mode and median** both in the *Severe* category. The strong dominance of higher pain levels suggests a focus on more severe patients, possibly due to the clinical setting in which the data was collected.

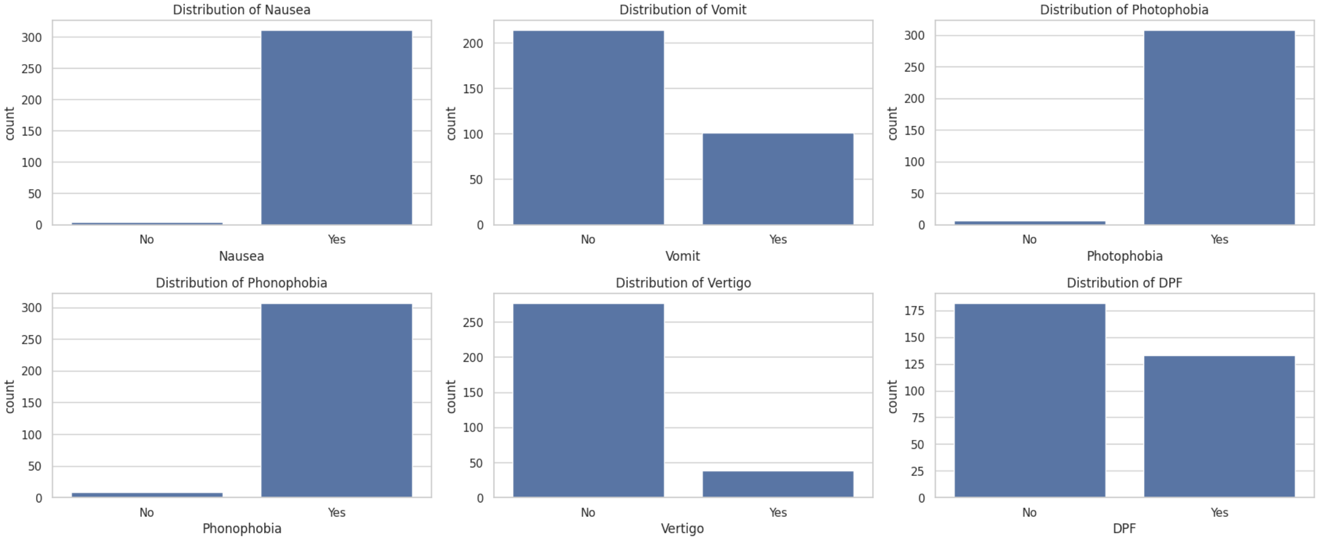


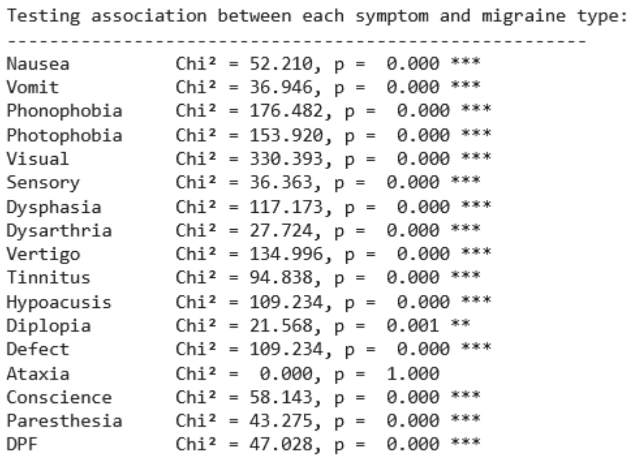
Figure 5:Count plots of selected binary variables (1 = Yes, 0 = No)

To understand the prevalence of common migraine symptoms, we initially examined a larger set of binary variables representing symptom presence (e.g., dysarthria, tinnitus, paresthesia). For clarity and clinical relevance, we selected six key variables for detailed analysis: **Nausea**, **Vomit**, **Photophobia**, **Phonophobia**, **Vertigo**, and **DPF** (Drop attacks / Postural dysfunction).

As shown in Figure X, most patients experience **Nausea (99%)**, **Photophobia (97.8%)**, and **Phonophobia (97.5%)**, confirming these symptoms as hallmarks of migraine. **Vomit** is less frequent, reported by approximately **32%**, while **Vertigo** is present in about **12%** of cases. Interestingly, **DPF** is nearly evenly distributed between 0 and 1, making it a more informative and balanced variable for further analysis.

The figure highlights both symptom dominance and variation across patients. These binary features offer insight into patient experience and can support more targeted analysis or symptom-based subgrouping in future work.

Bi-Variate Analysis of Target Variable with other variables



```
Testing association between each symptom and migraine type:
--------------------------------------------------------
Nausea          Chi² = 52.210, p =  0.000 ***
Vomit           Chi² = 36.946, p =  0.000 ***
Phonophobia     Chi² = 176.482, p =  0.000 ***
Photophobia     Chi² = 153.920, p =  0.000 ***
Visual          Chi² = 330.393, p =  0.000 ***
Sensory         Chi² = 36.363, p =  0.000 ***
Dysphasia       Chi² = 117.173, p =  0.000 ***
Dysarthria      Chi² = 27.724, p =  0.000 ***
Vertigo         Chi² = 134.996, p =  0.000 ***
Tinnitus        Chi² = 94.838, p =  0.000 ***
Hypoacusis      Chi² = 109.234, p =  0.000 ***
Diplopia        Chi² = 21.568, p =  0.001 **
Defect          Chi² = 109.234, p =  0.000 ***
Ataxia          Chi² =  0.000, p =  1.000
Conscience      Chi² = 58.143, p =  0.000 ***
Paresthesia     Chi² = 43.275, p =  0.000 ***
DPF             Chi² = 47.028, p =  0.000 ***
```

Figure 6: Testing association between each symptom

We examined the relationship between individual symptoms and migraine classifications using the chi-square test of independence, a method suitable for assessing associations between categorical variables.

The results of chi-square tests conducted between each symptom and the migraine type. We found that 16 symptoms demonstrated a statistically significant association with the migraine type, indicating their potential relevance in differentiating between migraine subtypes. Only one symptom(Ataxia) was found to be statistically insignificant. These results provide a solid foundation for selecting features in classification modeling and exploratory symptom analysis.

Figure 7 illustrates a percentage-based heatmap showing the prevalence of each symptom across the different migraine types. Each cell represents the proportion of patients within a migraine category who experienced a particular symptom. Darker colors denote higher percentages. This visualization helps identify which symptoms are more characteristic of specific migraine types, revealing patterns that may support both clinical interpretation and algorithmic classification.

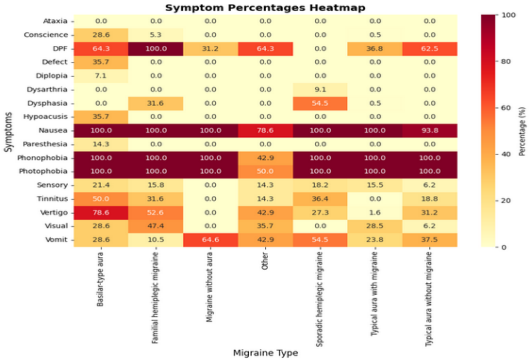


Figure 7: Symptom Percentage Heatmap

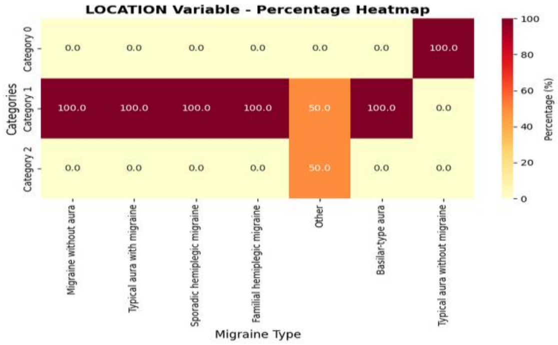


Figure 9: Location Variable; Percentage Heat map

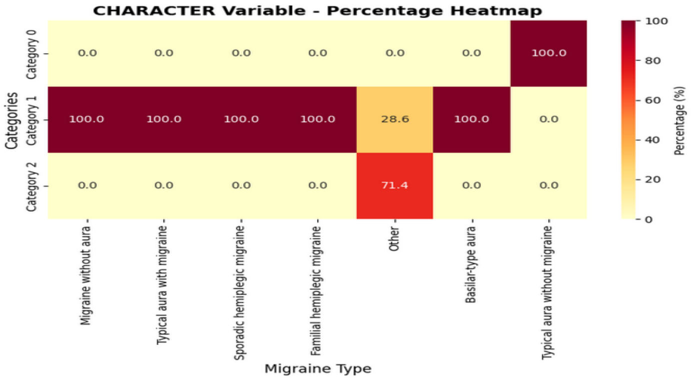


Figure 8: Character Variable; Percentage Heat map

Figures 8 and 9 illustrate the distribution of two key pain-related features—**location** and **character**—across different migraine types. Pain location (Unilateral, Bilateral, Frontal, or Temporal) shows notable variation among migraine categories, with **Unilateral pain being more prominent in types like Migraine without aura**, aligning with established clinical patterns. Similarly, the nature of pain (Throbbing, Pressing, Sharp, or Dull) highlights that **Throbbing pain is the most commonly reported character** across most migraine types, while other types such as Sharp or Dull appear more selectively. Together, these visualizations emphasize the diversity in how patients experience migraine pain and reinforce the importance of pain characteristics in refining migraine classification and diagnosis.

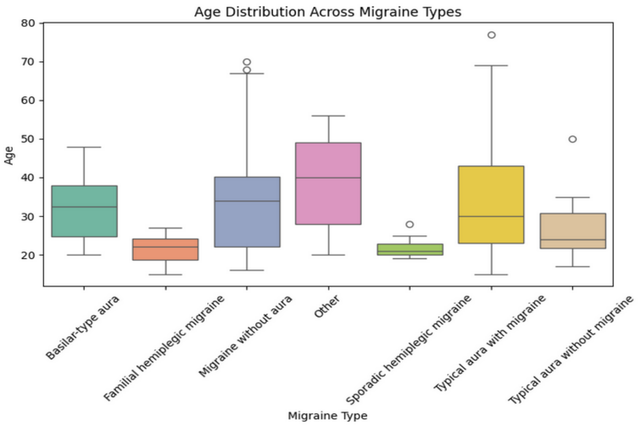


Figure 10: Age Distribution Across Migraine types

Figure 10 shows the age distribution of patients across the various migraine types. This distribution highlights that migraines tend to cluster around certain age groups depending on the type. For instance, some types may have a higher incidence in younger adults, while others are more evenly distributed. Identifying these age trends is important for epidemiological insights and may aid in age-specific diagnostic strategies.

Figure 11 displays the **distribution of attack frequency** across migraine types using a violin plot. Frequency refers to the number of migraine episodes experienced by a patient within a given time frame.

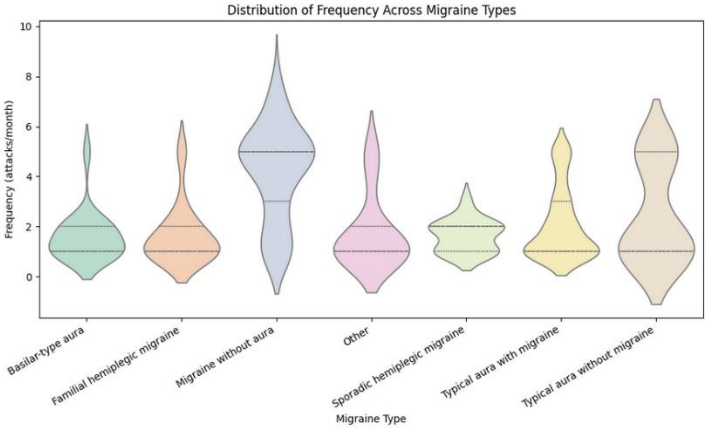
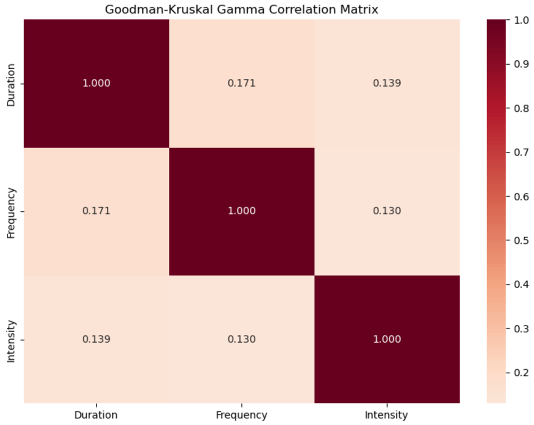


Figure 11: Distribution of frequency across Migraine types

The violin plot highlights the spread and density of frequency values for each migraine subtype. For example, *Migraine without aura* and *Typical aura with migraine* show a wide and dense distribution, suggesting that patients in these groups tend to experience a higher and more variable number of attacks. On the other hand, types like *Basilar-type aura* and *Familial hemiplegic migraine* exhibit narrower distributions, possibly reflecting a smaller, more homogeneous sample or fewer frequent episodes.

Multivariate Analysis

1. Goodman-Kruskal Gamma Correlation Matrix



This heatmap displays the Goodman-Kruskal gamma coefficients between three ordinal variables: Duration, Frequency, and Intensity of migraine symptoms. The diagonal shows perfect correlation (1.0) of each variable with itself, while off-diagonal values show pairwise associations.

Finding:

Figure 12:Goodman - Kruskal Gamma Correlation Matrix

- All three symptom characteristics show weak to moderate positive associations (coefficients between 0.13-0.17)
- The strongest association is between Duration and Frequency (γ = 0.171)

- Intensity shows slightly weaker associations with both Duration (γ = 0.139) and Frequency (γ = 0.130)

These moderate positive correlations suggest that longer migraine durations tend to occur with higher frequencies , More intense migraines tend to last longer and occur more frequently and No strong multicollinearity exists between these ordinal features.

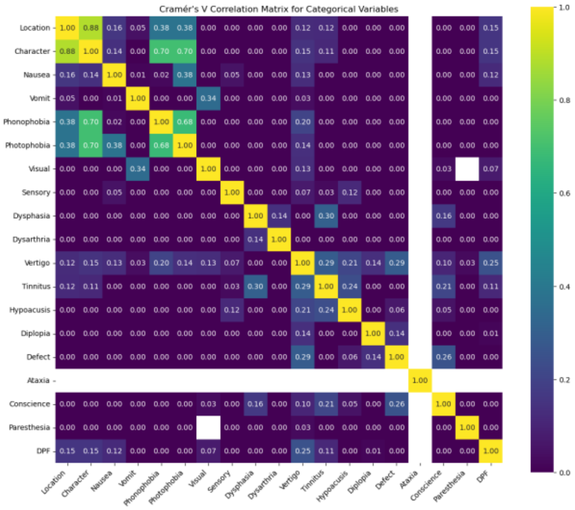


Figure 13: Cramer's V Correlation Matrix

2. Cramér's V Correlation Matrix

A large heatmap showing association strengths between all categorical variables using Cramér's V statistic, ranging from 0(no association) to 1 (perfect association).

Here:

- Several symptom pairs show strong associations (V > 0.5)
- Phonophobia and Photophobia are highly correlated.
- Visual symptoms show moderate associations with sensory symptoms
- Nausea and Vomit display moderate correlation (V ≈ 0.4).

Factor Analysis of Mixed Data (FAMD)

Given the mix of categorical and numeric variables in our migraine symptom dataset, Factor Analysis of Mixed Data (FAMD) is ideal. It merges the strengths of PCA and MCA to reduce dimensionality while preserving meaningful structure, offering a simplified view of symptom patterns.

Component Analysis and Explained Variance

To determine how many components are useful, we reviewed the eigenvalues and variance explained:

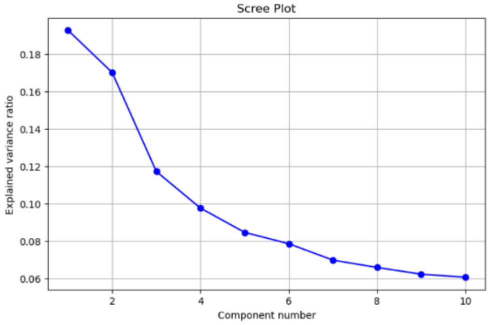


Figure 14: Scree Plot

Table 2:Table of eigenvalues and explained variance

Component	Eigenvalue	Explained Variance Ratio
1	35.86	0.200
2	31.52	0.175
3	21.40	0.119
4	16.36	0.091
5	15.04	0.084

The first component explains 20% of the variance, followed by 17.5% for the second. Together, the first four components account for nearly 58% of the total variance. The scree plot shows a notable drop after the third component, indicating the top three components are the most informative.

Patient Projection by Migraine Type

Figure 15 presents the FAMD projection of patients based on the first two components, with points colored by migraine type. A distinct cluster is observed for individuals with Typical aura without migraine, separated along the first component axis. Most other migraine types are concentrated near the origin, indicating similar underlying feature patterns. The wide spread of the "Other" category suggests potential outliers or heterogeneous symptom profiles.

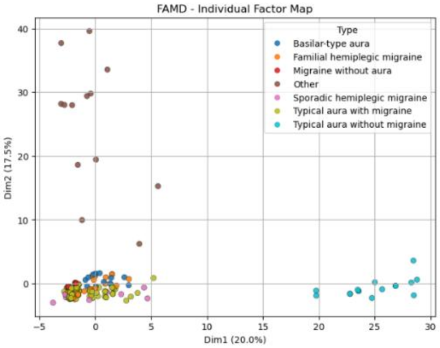


Figure 15: Individual Factor Map Plot

Variable Contributions to FAMD Dimensions

Next, we examined how individual symptoms contribute to the two key dimensions:

Table 3: Top Contributing Variables to Dimension 01 & 02

Dimension 01		Dimension 02	
Variable	Contribution	Variable	Contribution
Intensity	0.025926	Character	0.028207
Location	0.025875	Phonophobia	0.021807
Character	0.025872	Photophobia	0.020936
Visual	0.013893	Location	0.019950

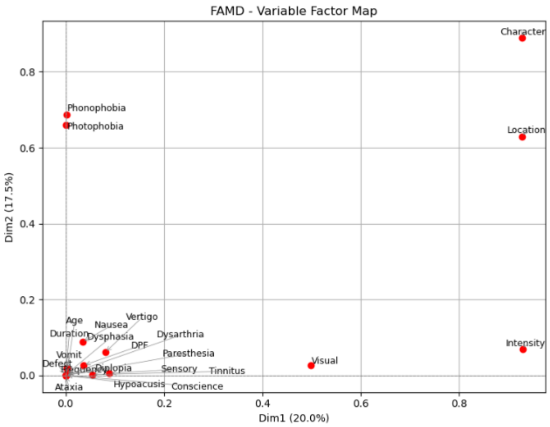


Figure 16: Variable Factor Map

The variable factor map and the corresponding table illustrate the top contributing migraine symptoms to the first two dimensions derived from the FAMD. Each point on the map represents a symptom or category, positioned based on its correlation with the first two dimensions. Variables farther from the origin have a stronger influence, while those near the center contribute less to the explained variance.

Symptoms like Intensity, Location, and Character contribute most to Dimension 1, suggesting this dimension captures core migraine traits. Dimension 2 is influenced by Phonophobia and Photophobia, pointing to sensory sensitivities. This reveals how symptoms vary across dimensions.

FAMD Biplot

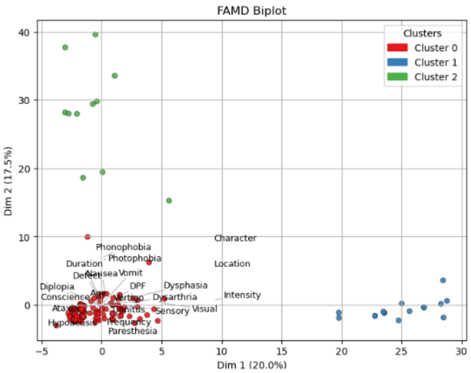
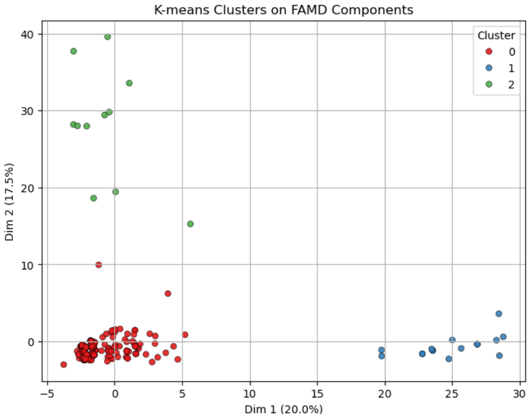


Figure 17: FAMD Biplot

Figure 17 displays the FAMD biplot combining patient projections and variable contributions, with individuals colored by cluster. Cluster 1 separates along Dimension 1 and is linked to symptoms like Visual and Intensity. Cluster 2 is more distinct along Dimension 2, associated with Photophobia and Phonophobia. Cluster 0 is near the origin, showing a more typical or mixed symptom pattern. This biplot helps us see how symptoms influence the clustering structure.

Clustering Interpretation and Patient Profiles

K-Means clustering on the FAMD components identified three distinct patient groups with a silhouette score of 0.908:



- **Cluster 0 (n = 288):** A large, balanced group (mean age ~ 31.5) with moderate intensity and common sensory symptoms, likely typical migraine cases.
- **Cluster 1 (n = 16)**: The youngest group (mean age~ 27.5), with low intensity and few neurological symptoms, possibly early or mild cases.
- **Cluster 2 (n = 11):** The oldest group (mean age ~ 39.9), with high intensity but lacking sensory symptoms, suggesting a more atypical profile.

These clusters reflect the diversity of migraine experiences, as captured through FAMD.

Figure 18: K-means Clustering

Analyzing Nonlinear Class Boundaries in Migraine Classification

The plotted decision boundaries clearly reveal that the separation among migraine subtypes is nonlinear. The curved and irregular boundaries distinguishing categories such as Hemiplegic, Typical Aura, Without Aura, and Other Subtypes indicate that the underlying data patterns are complex and not linearly separable. Consequently, linear approaches like Linear Discriminant Analysis (LDA) may fail to accurately represent these relationships. To capture the true structure and improve classification performance, nonlinear models—such as Support Vector Machines with RBF kernels, Random Forests, or Neural Networks—are more suitable for this takes.

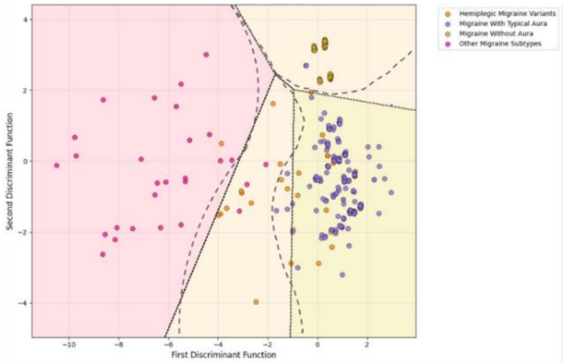


Figure 19: Decision Boundary Plot

SUGGESTIONS FOR A QUALITY ADVANCED ANALYSIS

Based on the exploratory analysis, several directions for advanced analysis are possible. A classification model could be developed to predict **migraine type** using patient symptoms, duration, and age. Techniques such as **Random Forests**, **XGBoost**, or **logistic regression** could be explored. To handle the **imbalanced response variable**, methods like **SMOTE** or **class weighting** should be considered.

For deeper insight, **feature importance** and **clustering** could be used to identify key symptom patterns or patient subgroups. Additionally, **statistical tests** such as chi-square or ANOVA could quantify the relationships between symptoms and migraine severity or type. These methods would enable a more robust understanding of the data and support data-driven decision-making in clinical contexts.

REFERENCES

https://vitalflux.com/kmeans-silhouette-score-explained-with-python-example/#google_vignette

https://www.neuroelectrics.com/blog/clustering-methods-in-exploratory-analysis

https://towardsdatascience.com/factor-analysis-of-mixed-data-5ad5ce98663c/

https://www.geeksforgeeks.org/data-visualization/what-is-univariate-bivariate-multivariate-analysis-in-data-visualisation/

https://www.mayoclinic.org/diseases-conditions/migraine-headache/symptoms-causes/syc-20360201

https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2025.1555215/full

https://www.scirp.org/journal/paperinformation?paperid=141994

https://www.ijisrt.com/assets/upload/files/IJISRT20OCT290.pdf

APPENDIX

Data Set Link:

- Migraine Symptom Dataset for Classification

Google Colab Links:

- https://colab.research.google.com/drive/1EUWR4IsZQJlfOtOQbCjtS3_c-BnpjHx3#scrollTo=ynPHy4h8yh6O
- https://colab.research.google.com/drive/1rIkPMqk3pU1gM_6x2PYztndDAVMymOgL?usp=sharing

Git Hub Link:

- https://github.com/Vayani-Kavindya/Project-01-ST-4052.git