**Exploratory Data Analysis: Myopia Study**

**Heyda I. Alvarado Alvelo**

**Dr. Ami Gates**

**August 27, 2024**

**Table of Contents**

[INTRODUCTION 3](#_Toc175849422)

[EXPLORATORY DATA ANALYSIS 4](#_Toc175849423)

[PREPROCESSING 9](#_Toc175849424)

[CONCLUSION 13](#_Toc175849425)

[References 15](#_Toc175849426)

[Appendix A 16](#_Toc175849427)

# INTRODUCTION

**Objectives and Rationale**

The primary objective of this analysis is to perform a comprehensive Exploratory Data Analysis (EDA) on a dataset related to myopia, with the aim of answering specific questions about the relationships between various factors and the development of myopia. We are particularly interested in determining whether there is a significant correlation between myopia and genetic factors, such as parental myopia, and whether age plays a role in the likelihood of developing this condition. Additionally, we aim to assess the balance of the dataset to determine its suitability for building predictive models, and to explore whether classification or regression models would be more appropriate for this data. To achieve these objectives, we have chosen to employ Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). PCA is used to identify and reduce the dimensionality of the dataset by focusing on the most influential features, while t-SNE helps in capturing and visualizing non-linear relationships that may exist within the data. These methods are highly relevant for this analysis as they allow us to uncover underlying patterns that may not be immediately apparent, providing a strong foundation for subsequent modeling efforts.

**Problem Domain**

Myopia, commonly known as nearsightedness, is a significant and growing public health concern. The condition affects an individual's ability to see distant objects clearly, and its prevalence is on the rise globally. According to the Myopia Institute, it is projected that by 2050, approximately 50% of the world's population will be affected by myopia. This increase has been linked to both genetic predispositions and environmental factors, such as extended periods of near-work activities, reduced time spent outdoors, and increased screen time (2024). The dataset used in this analysis includes a variety of demographic, genetic, and lifestyle variables, making it an ideal candidate for studying the multifactorial causes of myopia (Braglia). Understanding these factors is crucial for developing predictive models that can identify individuals at risk for myopia and inform targeted prevention strategies. By exploring the relationships within this data, we hope to contribute to broader efforts in mitigating the impact of this rapidly growing eye condition.

# EXPLORATORY DATA ANALYSIS

**Data**

The Myopia Study dataset (Braglia) is a valuable resource for understanding the development of myopia in children, gathered between 1990 and 1995. The dataset contains 18 columns and 618 rows, each representing a subject who was part of this longitudinal study.

This dataset, obtained from GitHub, includes a wide range of variables that provide insight into both biological and environmental influences on eye health. These variables encompass demographic information such as age and gender, specific eye measurements like Spherical Equivalent Refraction (SPHEQ) and Axial Length (AL), as well as lifestyle factors, including the number of hours spent on activities like reading, sports, and screen time. Additionally, the dataset includes information about parental myopia, which could be a significant factor in understanding the genetic predisposition to myopia.

**Explanation of Variables**

The dataset includes a variety of variables that can be categorized into three main groups: demographic information, eye health metrics, and lifestyle factors. See Figure 1 for the first five rows of the dataset.

1. **Demographic Information:**
   * **ID:** Unique identifier for each subject
   * **STUDYYEAR:** The year each subject entered the study.
   * **AGE:** Age of the subject at their first visit, recorded in years.
   * **GENDER:** Gender of the subject coded as 0 for male and 1 for female.
2. **Eye Health Metrics:**
   * **MYOPIC:** Indicates whether the subject developed myopia within the first five years of follow-up (0 = No, 1 = Yes).
   * **SPHEQ:** Spherical Equivalent Refraction, a measure of the eye's focusing power, recorded in diopters.
   * **AL:** Axial Length, the length of the eye from front to back, measured in millimeters.
   * **ACD:** Anterior Chamber Depth, the length from front to back of the aqueous-containing space between the cornea and the iris, measured in millimeters.
   * **LT:** Lens Thickness, the length from front to back of the crystalline lens, measured in millimeters.
   * **VCD:** Vitreous Chamber Depth, the length from front to back of the aqueous-containing space in front of the retina, measured in millimeters.
3. **Lifestyle Factors:**
   * **SPORTHR:** Number of hours per week spent engaging in sports or outdoor activities.
   * **READHR:** Number of hours per week spent reading for pleasure.
   * **COMPHR:** Number of hours per week spent playing video/computer games or working on the computer.
   * **STUDYHR:** Number of hours per week spent reading or studying for school assignments.
   * **TVHR:** Number of hours per week spent watching television.
   * **DIOPTERHR:** Composite variable representing the total hours spent on near-work activities.
4. **Parental Myopia:**
   * **MOMMY:** Indicates whether the subject's mother is myopic (0 = No, 1 = Yes).
   * **DADMY:** Indicates whether the subject's father is myopic (0 = No, 1 = Yes).

**Exploration**

To begin exploring the data, we leveraged several functions in Python to review the structure of the dataset. Initially, by using the **.shape** method (Figure 2), we determined that the dataset comprises 618 rows and 17 columns (after data cleaning), which gave us an overview of the dataset’s dimensions and the datatypes of each variable. To streamline the exploration process, we categorized the variables into numerical and categorical groups, defined as **num\_myopia** and **cat\_myopia**, respectively. This categorization allowed us to perform more targeted analysis and produce cleaner, more relevant results (Figure 2).

We then conducted a summary of the numerical variables using the **.describe()** function, which provided key statistics such as the mean, standard deviation, and percentiles for each numerical feature. The summary statistics revealed several important insights (Figure 2):

* **Central Tendency:** The mean values for variables like AGE, SPHEQ, and AL give us an understanding of the typical values within the dataset. For instance, the average age of subjects is approximately 6.30 years, and the average Spherical Equivalent Refraction (SPHEQ) is .80 diopters.
* **Dispersion:** The standard deviation for variables like SPORTHR and TVHR indicates significant variability in the number of hours spent on sports and watching TV, reflecting diverse activity levels among the subjects. For example, SPORTHR has a standard deviation of approximately 7.97 hours, suggesting that while some children spend considerable time in sports, others may engage in very little or none.
* **Ranges:** The min and max values for each variable indicate the breadth of the data. Notably, SPHEQ ranges from -0.69 to 4.37 diopters, and VCD (Vitreous Chamber Depth) spans from 13.38 mm to 17.30 mm, highlighting the diversity in eye measurements within the population.

In addition to numerical exploration, we explored the distribution of categorical variables using a **for loop** (Figure 3) to iterate over each category and print its value counts. The distribution of the MYOPIC variable, for example, showed a significant imbalance, with 537 non-myopic individuals compared to 81 myopic individuals. On the other hand, the GENDER variable is nearly balanced, with 316 males and 302 females, providing a fair representation across genders. The parental myopia variables, MOMMY and DADDY, are also relatively balanced, indicating that around half of the subjects had at least one myopic parent.

**Visualizations**

The exploratory data analysis continued with a series of visualizations to gain deeper insights into the dataset. Below is a summary of the key findings from these visualizations:

**1. Bar Graph by STUDYYEAR:** The bar graph (Figure 4) for the STUDYYEAR variable shows that the number of observations is relatively consistent across the years, with a peak in 1990. The dataset includes data from 1990 to 1995, with the highest number of entries in 1990 and the lowest in 1995.

**2. Bar Graph by AGE:** The age distribution seen in Figure 5, is highly skewed, with many subjects being 6 years old. This age group dominates the dataset, with more than 400 observations, while other ages (ranging from 5 to 9 years) are underrepresented. This skewness in the age distribution may affect the generalizability of the results, as most of the data reflects the characteristics of 6-year-olds.

**3. Histogram SPHEQ:** The distribution plot for SPHEQ in Figure 6 shows a clear concentration around values close to 1.0 diopters, with a sharp decline in frequency as the values increase or decrease from this point. The distribution suggests that most subjects have a mild refractive error. The distribution is slightly right-skewed, indicating a small number of subjects with higher positive refractive errors.

**4. Probability Density Function of SPHEQ:** The probability density function (PDF) for SPHEQ (Figure 7) confirms the observations from the histogram, with a peak around 1.0 diopters. The PDF suggests that most of the subjects have a refractive error close to this value, with the likelihood of encountering more extreme values decreasing as the refractive error moves away from the mean.

**5. Correlation Plot Using Pairplot in Relation to SPHEQ:** The **pairplot** (Figure 8) provides a detailed view of the relationships between SPHEQ and other numerical variables. The plots suggest weak to moderate correlations between SPHEQ and other variables such as AL (Axial Length) and VCD (Vitreous Chamber Depth). However, the overall scatter indicates significant variability, suggesting that SPHEQ is influenced by multiple factors.

**6. Correlation Matrix:** Our initial correlation analysis focused on the relationship between SPHEQ and other variables, but it revealed no significant correlations. Consequently, we expanded the analysis to include a correlation matrix for all variables, which provided a broader understanding of the dataset (Figure 9). This approach uncovered strong correlations between certain variables, such as the high correlation between AL (Axial Length) and VCD (Vitreous Chamber Depth) at 0.94, and between DIOPTERHR (near-work activities) and both READHR and STUDYHR, with correlations of 0.70 and 0.62, respectively.

# PREPROCESSING

**Data Cleaning**

The initial phase of data cleaning involved a series of exploratory checks to understand the structure and quality of the dataset. We began by inspecting the dataset's first five rows using the **myopia.head()** function (Figure 1). This step provided a preliminary visual assessment of the data, allowing us to observe the types of values present and assess any immediate inconsistencies or anomalies. Following this, we employed the **myopia.info()** function to obtain comprehensive information about the dataset. This function was instrumental in confirming the data types of each column, identifying the presence or absence of null values, and verifying the names and number of columns, as well as the overall dimensions of the dataset.

One of the key data cleaning actions taken was the renaming of the column **"DADMY"** to **"DADDY."** This step was crucial because consistent and accurate column naming is essential for clear communication and analysis. A typo in a column name can lead to confusion, misinterpretation of data, and errors in subsequent analysis. By using the rename function, we ensured that the dataset was correctly labeled, thereby enhancing the dataset's clarity and usability.

Next, we conducted a thorough examination for common data quality issues, such as null values, whitespace, and duplicate entries. Detecting and addressing these issues is a critical aspect of data cleaning, as they can significantly impact the validity and reliability of the analysis. In our dataset, there were no null values, extraneous whitespace, or duplicate records, which allowed us to proceed with confidence that the data was complete and clean.

Finally, we decided to remove the "ID" column from the dataset. The "ID" column, typically serving as a unique identifier for each record, was deemed irrelevant for the analysis since it does not contribute any meaningful information related to the study's objectives. Removing irrelevant columns is an important data cleaning step, as it reduces unnecessary complexity and ensures that only pertinent data is included in the analysis. This not only simplifies the dataset but also improves the performance of any subsequent data processing and modeling tasks. See Figure 10 for the Data Cleaning steps.

**Data Transformation**

Before conducting the PCA and t-SNE analyses, we focused on transforming the numerical features of the dataset to ensure they were appropriately prepared for these data reduction techniques. Using the scikit-learn library, we applied both normalization and scaling (Figure 11). Normalization involved adjusting the numerical features so that they each ranged between 0 and 1, thereby ensuring that all features contributed equally to the analysis. This step was crucial for preventing any one feature from disproportionately influencing the results due to differences in scale (*MinMaxScaler*).

In addition to normalization, we employed **StandardScaler** from scikit-learn to scale the data, setting each feature to have a mean of 0 and a standard deviation of 1. Scaling is particularly important in PCA, which assumes that the data is centered around the origin and that all features have unit variance. By normalizing and scaling the data, we ensured that our subsequent PCA and t-SNE analyses were both accurate and reliable, allowing us to uncover meaningful patterns within the dataset (*Importance of feature scaling*).

**Data Reduction**

To reduce the dimensionality of our dataset and gain insights into its structure, we performed Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), using the scikit-learn library for both techniques. Please see Figure 12 and Figure 13 for the code.

PCA is a linear dimensionality reduction technique that transforms the data into a set of orthogonal components, called principal components, which capture the maximum variance in the data (*PCA*). By projecting the data onto these principal components, PCA helps to simplify the dataset while retaining as much information as possible. In our analysis, we selected the first two principal components for visualization, which allowed us to observe the distribution of data points across these dimensions. However, as shown in our PCA scatter plot in Figure 14, the results confirmed that the dataset did not exhibit a clear linear relationship between the features, as the two classes (myopic and non-myopic) were largely overlapping and concentrated around the center, indicating that PCA could not distinctly separate the classes.

Recognizing the limitations of PCA in handling non-linear relationships, we then applied t-SNE, a non-linear dimensionality reduction technique designed to preserve the local structure of the data. t-SNE is particularly effective at revealing complex patterns by mapping high-dimensional data into a lower-dimensional space, typically 2D or 3D, where it can be more easily visualized (*Tsne*). After performing t-SNE and visualizing the results in Figure 15, we observed some potential clustering, suggesting that t-SNE could capture more subtle relationships in the data that PCA missed. However, like PCA, the t-SNE plot did not reveal a clear separation between the myopic and non-myopic classes, though it hinted at areas where clusters might be present.

# CONCLUSION

**Summary**

The analysis of the Myopia Study dataset yielded several important observations, particularly regarding the structure and balance of the data. Notably, the dataset is unbalanced, with a significant skew in both the MYOPIC variable—where non-myopic individuals far outnumber myopic ones—and in the age distribution, where most subjects are 6 years old. These imbalances could impact the development and accuracy of predictive models. Despite applying Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), we did not observe strong correlations between the key variables of interest, such as SPHEQ (Spherical Equivalent Refraction) or the MYOPIC target variable, and other features like parental myopia or age. Additionally, neither PCA nor t-SNE provided clear separation or clustering of the myopic and non-myopic classes, suggesting that these techniques alone may not be sufficient for identifying meaningful patterns in this dataset.

**Limitations**

The limitations of this analysis largely stem from the characteristics of the dataset and the techniques employed. The unbalanced nature of the data, particularly in the MYOPIC and AGE variables, likely limited the effectiveness of our exploratory techniques. The skewed distribution in these areas could lead to biased or unreliable results, particularly when attempting to build predictive models. Furthermore, while PCA and t-SNE are valuable tools for dimensionality reduction and visualization, they did not produce significant insights in this case. The lack of clear correlations between SPHEQ or the MYOPIC target variable and other factors suggests that more complex relationships may exist, which these linear and non-linear techniques were unable to capture fully.

**Improvement Areas**

To improve the analysis and potential modeling efforts, future work should consider balancing the dataset, particularly in terms of the MYOPIC and AGE variables. Techniques such as oversampling, undersampling, or synthetic data generation could help address these imbalances. Additionally, exploring alternative or more advanced dimensionality reduction techniques, such as Uniform Manifold Approximation and Projection (UMAP) or deep learning approaches, may provide better insights into the non-linear relationships within the data. Finally, incorporating additional features or external data sources could help to better capture the complexities of myopia development and lead to more accurate and robust predictive models.

# References

Awan, A. A. (2023, March 30). *Introduction to T-SNE: Nonlinear dimensionality reduction and Data Visualization*. DataCamp. <https://www.datacamp.com/tutorial/introduction-t-sne>

Braglia, L. (n.d.). *Aplore3/rawdata/myopia at master · lbraglia/APLORE3*. GitHub. <https://github.com/lbraglia/aplore3/tree/master/rawdata/MYOPIA>

*Importance of feature scaling*. scikit. (n.d.-a). <https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html#sphx-glr-auto-examples-preprocessing-plot-scaling-importance-py>

*MinMaxScaler*. scikit. (n.d.-b). <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html#sklearn.preprocessing.MinMaxScaler>

*Myopia - Myopia Institute*. Myopia Institute -. (2024, February 19). <https://myopiainstitute.org/myopia/>

*PCA*. scikit. (n.d.-c). <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#pca>

*Tsne*. scikit. (n.d.-d). <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>

# Appendix A

**Figure 1**: First Five Rows of the Myopia Dataset

A screenshot of a computer screen

Description automatically generated

**Figure 2:** Data Exploration – Dataset dimensions, defining numerical and categorical features & Descriptive Statistics

A screenshot of a computer

Description automatically generated

**Figure 3:** for loop – Count of the Categorical Features

A screenshot of a computer

Description automatically generated

**Figure 4:** Bar Graph by STUDDYYEAR

A graph of blue bars

Description automatically generated with medium confidence

**Figure 5:** Bar graph by AGE **Figure 6:** SPHEQ Histogram

A graph of a number of people

Description automatically generatedA graph of a number of blue bars

Description automatically generated

**Figure 7:** Probability Density Function for SPHEQ

A blue line graph with numbers

Description automatically generated

**Figure 8:** Correlation Plot Using pairplot in Relation to SPHEQ

A group of blue and white graphs

Description automatically generated

**Figure 9:** Correlation Matrix (ALL VARIABLES)

A screen shot of a chart

Description automatically generated

**Figure 10:** Data Cleaning Steps Code

A screenshot of a computer program

Description automatically generated

**Figure 11:** Normalization and Scaling of Numerical Features

A screen shot of a computer program

Description automatically generated

**Figure 12:** PCA Code

A screen shot of a computer program

Description automatically generated

**Figure 13:** t-SNE Code

A screen shot of a computer code

Description automatically generated

**Figure 14:** PCA Results Plot

A red and blue dots

Description automatically generated

**Figure 15:** t-SNE Results Plot

A red and blue dots

Description automatically generated