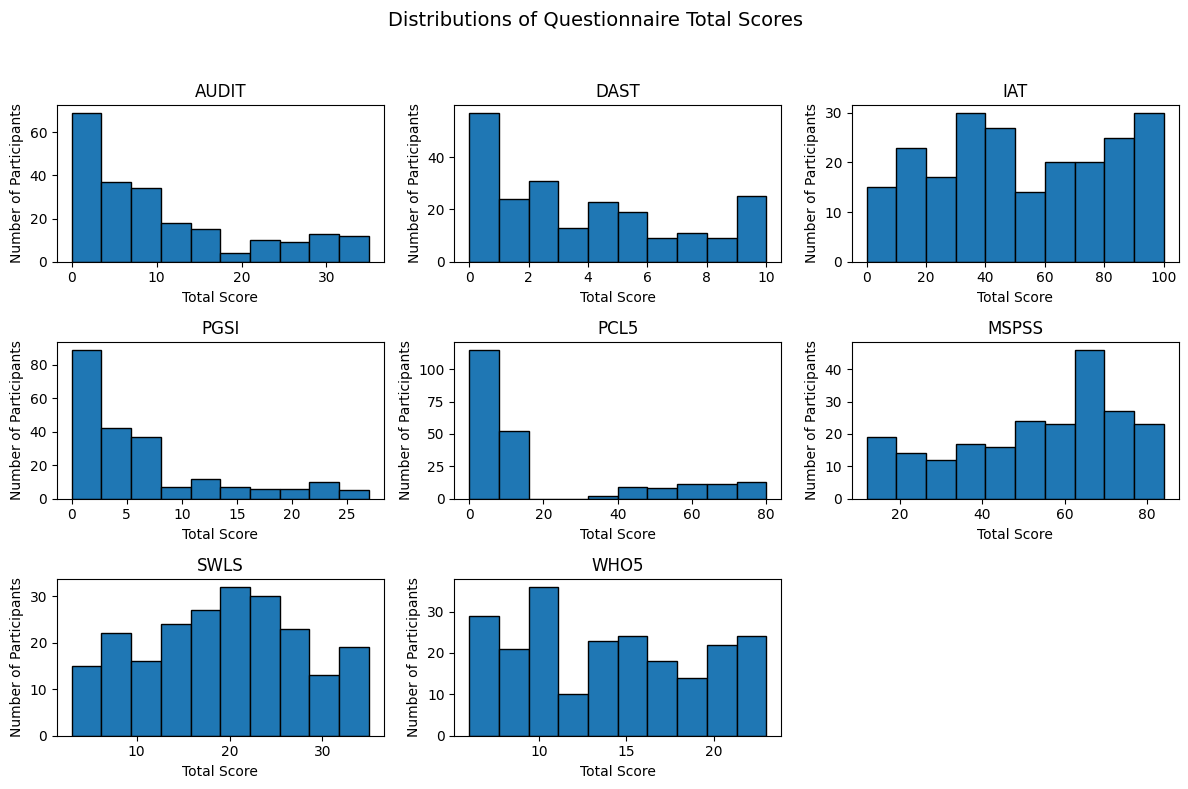
# **E-Health Project Report Group 13**

## I. Data Preparation and Exploratory Data Analysis

The dataset used in this project (dataset\_project\_eHealth20252026) originates from the eHealth study and includes both personal information (age, gender, education, marital status, income) and psychological questionnaire scores. In addition, we used the codebook to interpret the data correctly(questionnaire\_codebook\_eHealth20252026).

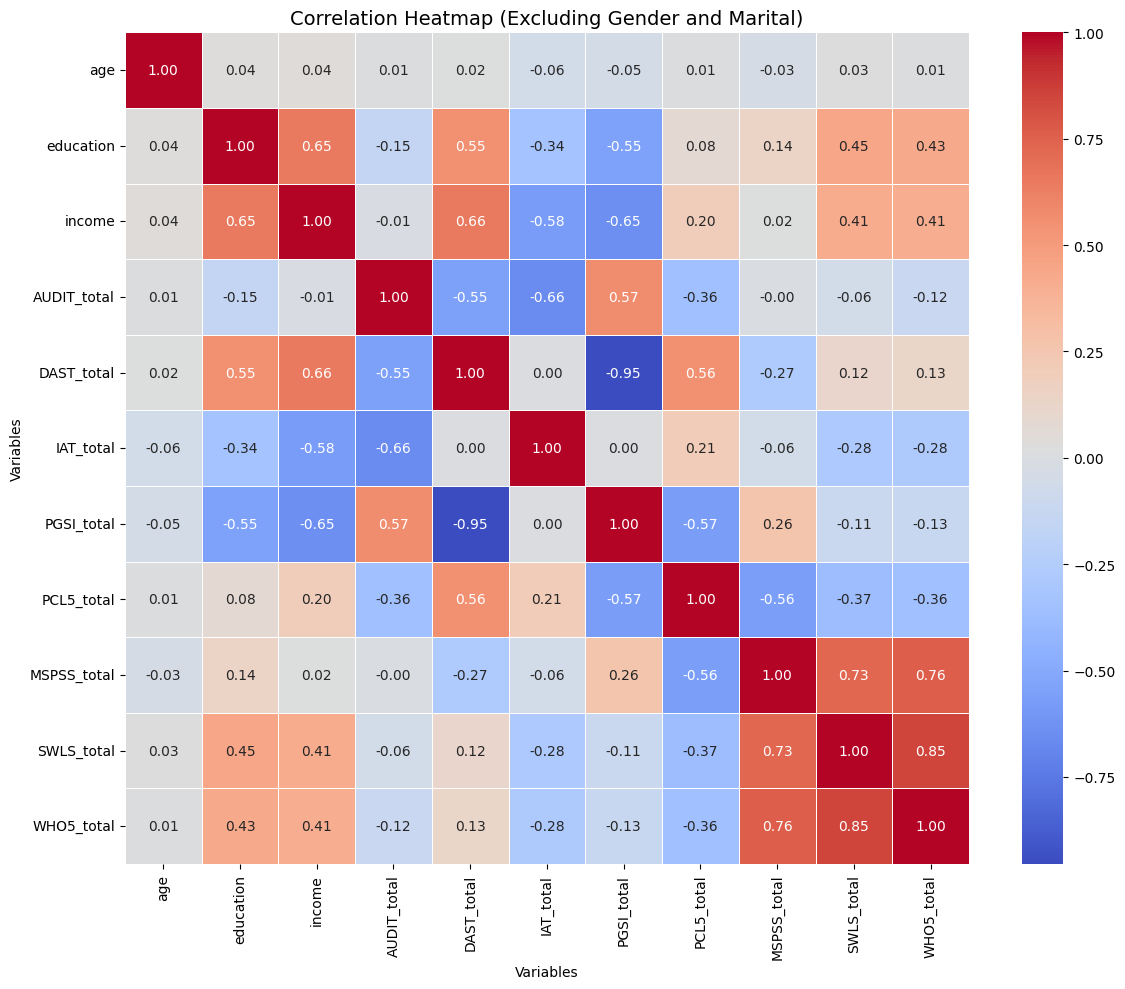
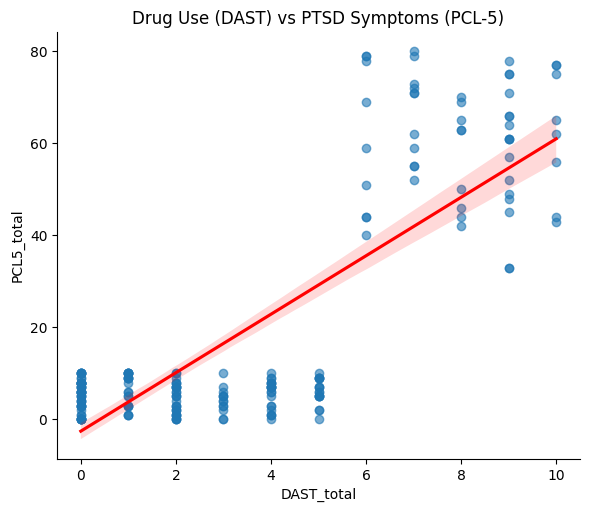
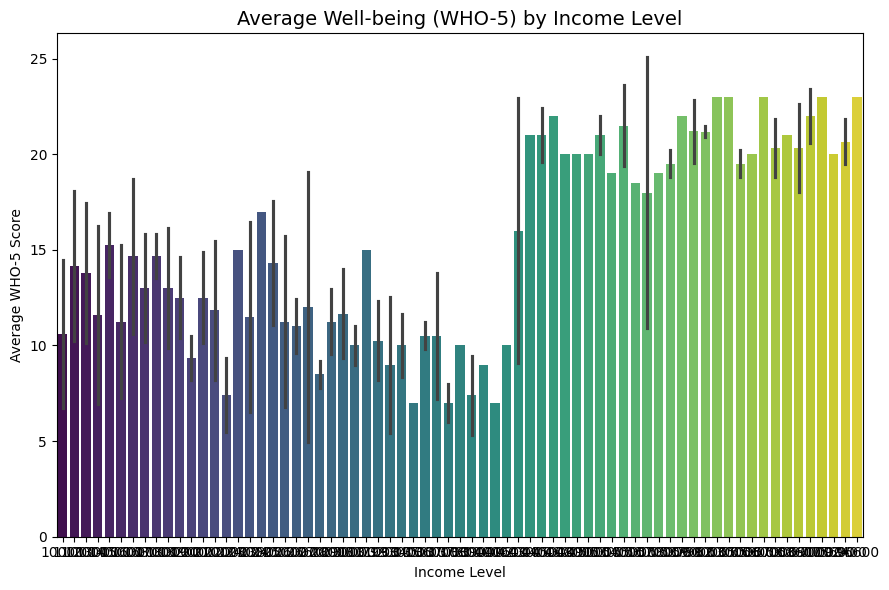
The dataset required cleaning due to inconsistencies, missing values. First, we replaced NaN/empty string values with either median (numerical cols)/most frequent values (categorical cols). Then we searched for outliers in the personal columns. We performed an outlier analysis on all questionnaire variables using z-scores (|z| > 3). No statistical outliers were detected in any of the 91 numeric questionnaire columns, and no participant had an abnormal response profile. Although no extreme values were found, the distributions remained highly skewed and non-normal, which justified why we later use RobustScaler for normalization.

*Figure 1: Histograms of totals*

Total scores were computed for each questionnaire, adding 8 new columns to the original dataset, one for each questionnaire, meaning the total score. (Figure 1)

To prepare the dataset for clustering, only questionnaire totals were normalized, so that they would be comparable. We used RobustScaler to normalize the questionnaire totals because the data is skewed and contains outliers, and methods like StandardScaler or MinMaxScaler are not that useful in these conditions. RobustScaler works with the median and IQR instead of the mean and standard deviation, so extreme values don’t shift the scale. This makes the transformed features much more stable for PCA and clustering, and gives a more reliable representation of the underlying patterns in the dataset.

EDA was performed to understand distributions, detect anomalies, and verify assumptions. We started with a heatmap, to understand a priori the possible correlations between addictions (Figure 2). Next we compared addictions two by two. Figure 3 demonstrates a strong positive relationship between drug use and PTSD symptoms, indicating that greater substance use is associated with more severe trauma-related distress. Figure 4 shows, as income level increases, the average WHO-5 score tends to rise as well. This means participants with higher income generally report better emotional well-being and life satisfaction. The relationship isn’t perfectly linear (some fluctuations), but the overall trend is upward.



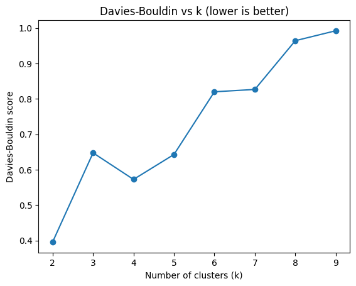
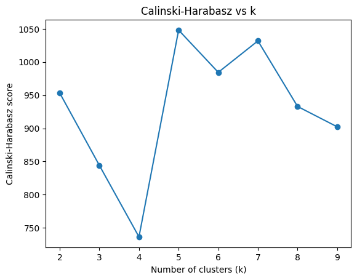
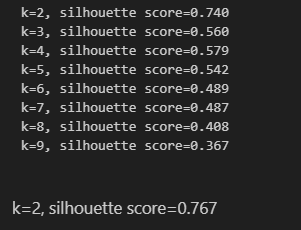
*Figure 2: Heatmap of totals*

*Figure 4: Correlation alcohol/income*

*Figure 3: Correlation drug use/ PTSD*

## II. Personas Creation

The last step of the data preparation was the normalization using RobustScaler. That transformed the data into a dataframe of normalized totals, together with the personal columns (gender, age, education, marital status, income). The personal columns were not scaled with Robust, but were used later when interpreting the clusters.

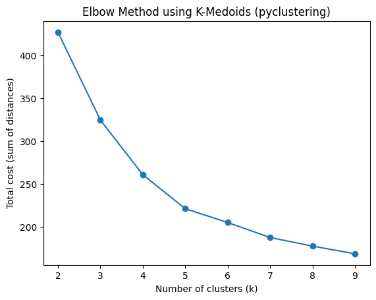
 Among the graph methods used to establish the final number of personas (k = nr of clusters) we enumerate elbow method (Fig. 5), silhouette score (Fig. 6), Calinski-Harabasz (Fig. 7), and Davies-Bouldin (Fig. 8). The internal validity indices showed mixed but informative patterns across different values of k. The Calinski–Harabasz (CH) index was relatively high for k = 2 and increased again for solutions with 5–8 clusters, with local peaks at k = 5 and k = 8, suggesting that more granular partitions can improve between-cluster separation. In contrast, the Davies–Bouldin (DB) index was lowest for k = 2 and increased for larger k, with comparatively favorable values around k = 4–5. Taken together with the elbow plot, we focused on a 4–5 cluster solution and ultimately retained the 4-cluster model.

*Figure 5: Elbow Method results*

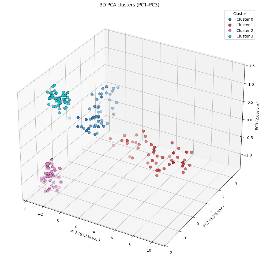
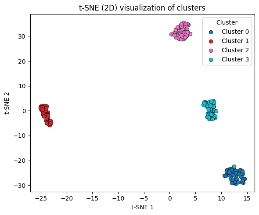
*Figure 6: Silhouette scores*

*Figure 7: Calinski-Harabasz vs k*

*Figure 8: Davies-Bouldin vs k*

Inside clustering, we chose 5 personal columns and 8 robust totals from the questionnaires. We used K-Medoids because our dataset contains non-normal distributions and variables measured on ordinal or bounded scales (psychological questionnaire totals). Whilst K-Means is sensitive to outliers and relies on Euclidean distance around a mean, K-Medoids is more robust because it chooses actual data points as cluster centers and minimizes distances without being distorted by extreme values.

A diagram of clusters with red blue and white dots

AI-generated content may be incorrect.A graph of a cluster

AI-generated content may be incorrect.The heatmap (Fig. 9) shows the mean robust-scaled scores for each cluster. Positive values indicate higher-than-median scores, while negative values indicate lower ones. Cluster 2 has clearly elevated AUDIT and PGSI scores - higher alcohol use and gambling risk. Cluster 3 shows very high PCL-5 and very low social support - strong PTSD symptoms. Cluster 1 displays the highest life satisfaction and well-being, with generally low risk levels. Cluster 0 is mainly characterized by elevated internet-use scores (IAT) while remaining average on most other scales.

*Figure 12: t-SNE in 2D*

*Figure 11: PCA in 2D*

*Figure 10:PCA in 3D*

*Figure 9: Heatmap of clusters*

PCA was then applied—not for clustering, but to visualize the structure of the data in two and three dimensions (Fig. 10, Fig. 11). By plotting the cluster assignments on the PCA components, we assessed how clearly the groups separated and how interpretable they were. PCA also provided insight into the underlying dimensions of variation in the data, which later supported the creation of meaningful personas. We also explored non-linear embeddings using t-SNE(Fig. 12). The two-dimensional t-SNE map showed a very clear separation between the four k-medoids clusters.

After generating the cluster assignments, we tested whether the clusters differed on the psychological questionnaire scores and demographic variables. For continuous scales (AUDIT, DAST, IAT, PGSI, PCL-5, MSPSS, SWLS, WHO-5), we used the Kruskal–Wallis test to detect overall differences across clusters. Whenever a variable showed a significant effect, we performed pairwise Mann–Whitney U tests with Bonferroni correction to identify which specific clusters differed from each other. This allowed us to confirm that the clusters represent statistically distinct groups.

A table with numbers and letters

AI-generated content may be incorrect.Categorical variables (gender, education level, marital status, income categories) were tested separately to examine if their distributions varied across clusters. We used either a Chi-square test or Fisher’s Exact Test to see if demographic proportions differed significantly across groups. The markers (C0=\**, C1=#, C2=$, C3=%)* next to each value show which clusters differ significantly from one another. Each symbol corresponds to a specific cluster and each cluster represents a persona.

*Figure 13: Personas table*

The four personas derived from the clustering reflect distinct psychological and demographic profiles. Each one of them summarizes the typical characteristics of a group: their background, frustrations, motivations, coping behaviors, and how the game should interact with them. The information comes directly from the significant psychological variables, demographic patterns, and risk profiles identified during clustering. All of this information was gathered under a persona card (Fig. 14).

A screenshot of a web page

AI-generated content may be incorrect.Cluster 0 represents overwhelmed adults with lower income and lower-to-medium education, showing moderate alcohol, drug, and internet-use scores, mid-level well-being, and meaningful reliance on social support. Cluster 1 captures socially active individuals—mostly men with lower education—who present the lowest internet-addiction scores but elevated alcohol use, gambling risk, and reduced life satisfaction. Cluster 2 reflects highly educated, higher-income adults who are strongly engaged online; they show high internet-use scores but overall good well-being, strong social support, and balanced psychological functioning. Cluster 3 includes emotionally resilient adults with moderate substance-use levels and higher trauma exposure, yet they maintain strong well-being and solid support networks. Together, these personas help translate statistical patterns into meaningful user profiles for guiding game design and intervention strategies.

*Figure 14: Persona card of cluster 0*

## III. Game Design

- Game Concept: theme, story summary, flow.

- Target Audience & Use Cases: define player types and contexts.

- Personalization Techniques: how personas affect the game.

- Mechanics & Technical Notes: UI flow, inputs, GitHub link.