Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Customer Segmentation of XYZ Sports Company

<Group 11>

Henrique Seganfredo, number:20230474

Ramzi Alayass , number: 20210705

Ehis Jegbefumen , number: 20221015

January, 2024

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. Data Exploration

Introduction

This report describes the customer segmentation process of a sports company, Alongside a cluster analysis, provides an explanation for each cluster identified, offering additional insights for marketing interpretation.

The XYZ Sports Company is a long-standing fitness facility dedicated to serving the community for numerous years. Customer segmentation can categorize individuals based on demographic, geographic, psychographic, and behavioral traits. These segments comprise customers sharing key similarities, enabling targeted marketing strategies in the future. Additionally, evaluating the Lifetime Value of customers within each segment holds relevance.

Initial Analysis

The XYZ Sports Company aims to acquire new customers by leveraging insights from its existing customer base. This initiative involves extracting comprehensive knowledge from available data. Understanding customer demographics and identifying primary customer segments is crucial for optimizing marketing investments. The goal is to shift from mass marketing strategies to tailored campaigns, encompassing cross-selling and upselling opportunities.

Our task is to provide detailed information on these customer groups, outlining their distinctive characteristics and potential product preferences. The project's implementation will utilise Python, with flexibility regarding the libraries employed. The dataset, derived from the company's ERP system, spans from June 2014 to October 2019, encompassing various customer-related attributes such as demographics, enrollment details, activity participation, visit frequencies, and enrollment statuses. With 14942 records and 31 features, it includes information on user IDs, age, gender, income, enrollment periods, activities, visit counts, renewals, references, and dropout statuses.

The dataset encompassed diverse features with distinct characteristics, necessitating varied treatment strategies in the data mining process. These distinctions arise from the nature of the features, ranging from numeric variables requiring imputation of missing values and outlier handling to categorical variables demanding binary encoding and unique value exploration. The heterogeneity of feature types underscores the importance of tailored preprocessing techniques to ensure meaningful insights and accurate model performance.

metric_features = "Age","Income", "DaysWithoutFrequency", "LifetimeValue","NumberOfFrequencies", "AttendedClasses","AllowedWeeklyVisitsBySLA", "AllowedNumberOfVisitsBySLA","RealNumberOfVisits","NumberOfRenewals", "NumberOfReferences".

non_metric_features="Gender","UseByTime","EnrollmentStart","EnrollmentFinish","LastPeriodStart" "LastPeriodFinish","DateLastVisit","AthleticsActivities","WaterActivities","FitnessActivities","DanceActivities","TeamActivities","RacketActivities","CombatActivities","NatureActivities","SpecialActivities", "OtherActivities","HasReferences","Dropout"

2.Data Preprocessing

Filling missing values

The initial analysis revealed missing values in several columns. The approach to handle missing values involves filling them with appropriate measures such as medians for numeric features and modes for non-numeric features. However, it's worth noting that the column "Income" was filled with medians, which may be reconsidered where filling with zeros could be more appropriate. Also, the "Income" feature, given its high number of NA values, should be handled carefully and could be considered for removal. Where AllowedWeeklyVisitsBySLA also a high number of missing values 535 under name nun-metric feature which also would be reconsidered. Thus except for income and AllowedWeeklyVisitsBySLA all other values have very few NAs so we are fine to forcibly push the medians and the modes.

Uniqueness Analysis/Value Count

The uniqueness analysis is conducted using the calculate_uniqueness function, which computes the percentage of unique values for each feature in the dataset. The resulting DataFrame (result_uniqueness) provides insights into the distribution of unique values across different features.

High Uniqueness Features: ID: All instances have unique values, making it a non-discriminatory feature. LifetimeValue: Shows moderate uniqueness at 37.93%. Moderate Uniqueness Features: EnrollmentStart, DateLastVisit: Exhibit moderate uniqueness at around 9%. Low Uniqueness Features: NatureActivities, DanceActivities: Have only one unique value, suggesting limited discriminatory power.

Binary Features: SpecialActivities, OtherActivities, CombatActivities, RacketActivities, TeamActivities, FitnessActivities, WaterActivities, AthleticsActivities, HasReferences, UseByTime, Gender, Dropout: These binary features have two unique values each and require careful handling during preprocessing. Recommendations: Feature Reduction/Dropping: NatureActivities, DanceActivities: With only one unique value, these features may be candidates for reduction or dropping, as they provide limited discriminatory information. Hence, understanding the uniqueness of features is crucial for effective preprocessing. The provided analysis highlights features with high, moderate, and low uniqueness, providing a foundation for further preprocessing steps. Careful consideration of binary features is recommended to ensure accurate model training.

The **value counts analysis** provides valuable insights into the distribution and patterns within the dataset. These findings are instrumental in guiding subsequent preprocessing steps and formulating targeted strategies for customer engagement, retention, and marketing. This analysis aims to unveil patterns, distributions, and potential insights into the dataset.

Key Findings: **Income** Distribution as a significant number of students (2123 instances) report zero income. This suggests a considerable portion of the dataset comprises individuals with no reported income, which may impact certain analyses and should be considered in future modeling. **Gender** Distribution: There are more females (1) than males (0) in the dataset, with 8931 instances of females and 6011 instances of males. Understanding this gender distribution is crucial for gender-specific analyses and targeted marketing strategies. **UseByTimeEnrollment** has the majority of

students (14238 instances) do not make use of the UseByTime enrollment option, indicating a low adoption rate for this particular enrollment method. Engagement in Activities: **DanceActivities** and **NatureActivities**: These features have no variation in values, with all instances having the same value. Further investigation is needed to understand if these features provide any meaningful information or can be omitted due to lack of variability. **OtherActivities**: A small number of instances (28) engage in other activities, which might need specific attention or categorization in subsequent analyses. **NumberofReferences** The dataset contains a low number of customers with references (297 instances), indicating that the majority of customers do not have references associated with them. This might be a point of interest for referral programs or understanding customer satisfaction and loyalty. **Dropout** Distribution: A significant number of students (11968 instances) are labeled as dropouts (1=true). Marketing strategies should be devised to re-engage and retain these students, as they represent a substantial portion of the dataset.

Date features manipulation/extraction

We made important changes to the date information. We converted dates related to enrollment and customer visits into a format that's easier for analysis. This allowed us to calculate useful time-based metrics like how long someone is enrolled ('EnrollmentDuration'), the duration of their last visit ('LastPeriodDuration'), and the time since their last visit ('TimeSinceLastVisit'). After doing this, we removed the original date columns from the dataset, making it more straightforward for further analysis. We also looked into how gender influences customer behavior. We found that there's not much difference in age and income between genders, but males tend to have higher long-term returns. Additionally, males show higher engagement metrics, like more frequent visits and longer durations as customers. Based on these findings, we suggested tailoring marketing strategies and engagement programs to specific genders and keeping a close eye on customer retention efforts. This information is crucial for our future modeling and decision-making in the project.

Outlier removal/Reduce noise data

We initially explored outlier removal using IQR limits but abandoned it due to excessive data loss (53% retention). Instead, we opted for a controlled approach, manually filtering outliers based on insights from data exploration. Specific limits were set for key features like 'Age,' 'Income,' 'DaysWithoutFrequency,' 'LifetimeValue,' 'NumberOfFrequencies,' 'AttendedClasses,' 'AllowedWeeklyVisitsBySLA,' 'RealNumberOfVisits,' and 'NumberOfReferences.' Notably, 'Age' was constrained between 0 and 87 to address potential inaccuracies.

After manual filtering, the resulting dataset ('df_filtered') retained 97.29% of the original rows, ensuring a balanced and representative dataset for subsequent analysis and modeling in our project. This controlled outlier removal contributes to refining data quality and sets the foundation for meaningful insights in later stages.

Feature engineering

In our efforts to enhance the dataset for our project, we carefully used insights from Exploratory Data Analysis (EDA) to guide our feature engineering decisions. A key focus was removing features highly correlated with others. Despite the notable correlation of 0.84 between "Income" and "Age," we intentionally opted not to eliminate the "Income" feature from our dataset. This decision stems

from the recognition of the inherent importance of income in a marketing context. While, "NumberOfRenewals" was removed because of its pronounced correlation (0.71) with "LifetimeValue" and its logical connection to customer loyalty.

Additionally, we examine features providing low information, leading to the removal of "NumberOfReferences" due to its minimal contribution beyond what was covered by the "HasReferences" feature. Features like "AllowedWeeklyVisitsBySLA" and "AllowedNumberOfVisitsBySLA" were considered irrelevant for our project's scope, pertaining to behavior-related specifics for specific customer types.

Considering clustering, we made thoughtful decisions about dummy (binary) features. "Gender" was evaluated for potential exclusion to prevent clustering polarization, recognizing that although males were a minority, they showed higher loyalty. "UseByTime" was identified as behavior-related but deemed beyond our project's focus, and "Dropout" was excluded as predicting dropout was not within our project's scope. "HasReferences" was considered too narrow for our project's objectives.

Features seemingly correlated with "RealNumberOfVisits," such as "NumberOfFrequencies" and "AttendedClasses," were retained for further evaluation to determine their impact on clustering. To proactively reduce dimensionality, we introduced the "ActivityEngagement" feature, consolidating various activities within the sports facility.

In summary, these strategic decisions reflect a meticulous approach to refining the dataset, ensuring alignment with our project's goals. The focus on reducing redundancy, eliminating irrelevant features, and considering clustering implications establishes a solid foundation for deriving meaningful insights in the subsequent phases of our data mining project.

Data Normalization

In the context of preparing our data for clustering methods, particularly k-means, and considering the removal of outliers, we have employed min-max scaling as part of the normalization process. Using the MinMaxScaler, we transformed our dataset, focusing on key metric features like 'Age, 'DaysWithoutFrequency,' 'LifetimeValue,' 'NumberOfFrequencies,' 'AttendedClasses,' 'RealNumberOfVisits,' 'EnrollmentDuration,' 'LastPeriodDuration,' 'TimeSinceLastVisit,' and the newly introduced 'ActivityEngagement.' This scaling ensures that all these features are proportionally represented within a common range (0 to 1), facilitating a more effective clustering process. The resulting scaled dataset, named 'df_scaled,' demonstrates a uniform scaling across the selected metrics, setting the stage for robust and accurate clustering analysis. This normalization step is vital for enhancing the comparability of features and optimizing the performance of clustering algorithms on our refined dataset.

Dimensionality Reduction/Principal Component Analysis

We addressed the polarization issue in the 'ActivityEngagement' feature by employing binary PCA (Principal Component Analysis). This technique allowed us to summarize various special activities into a set of principal components, enhancing the interpretability and efficiency of subsequent clustering analyses.Initially, we conducted a simulation with a high component count (n_components = 10) to explore the optimal number of components. The scree plot and explained variance analyses

revealed that retaining three principal components (PC=3) adequately explained over 80% of the variance, prompting us to proceed with this configuration.Following this determination, we performed PCA again with n_components = 3, generating three principal components named 'PCA_Activity_0,' 'PCA_Activity_1,' and 'PCA_Activity_2.' Interpretation of these components indicated significant contributions from activities such as 'WaterActivities,' 'FitnessActivities,' 'TeamActivities,' and 'CombatActivities,' while 'DanceActivities' and 'NatureActivities' had negligible impact due to constant values.

To integrate the newly derived principal components into our dataset, we updated the feature lists and dropped the original dummy activity features. However, an issue remained – our dataset was min-max scaled for metric features, but the PCA columns seemed to have different scales. Recognizing the potential impact on k-means clustering results, we applied MinMaxScaler to ensure uniform scaling for all features, promoting fair contributions to the clustering process. This comprehensive dimensionality reduction process involving binary PCA not only addressed the polarization issue in 'ActivityEngagement' but also optimized the dataset for subsequent clustering algorithms, laying the foundation for meaningful insights.

Ending Preprocessing

In the final steps of our preprocessing phase, we created two distinct datasets to facilitate subsequent clustering trials. The first dataset, referred to as 'df_without_activity,' excludes the 'ActivityEngagement' column, effectively removing the target variable. This dataset retains all other relevant features and serves as our reference for the subsequent analysis. To enhance the representation of special activities and accommodate the dimensionality reduction achieved through PCA, we constructed a list of metric features named 'metric_features_pca.' This list comprises all metric features from the original dataset, excluding 'ActivityEngagement' and incorporating the newly derived PCA components ('PCA_Activity_0,' 'PCA_Activity_1,' and 'PCA_Activity_2').

The 'df_without_activity[metric_features_pca]' dataset, derived from 'df_without_activity,' now serves as our primary data source for subsequent clustering trials. This dataset contains no categorical data, exclusively holding metric features, and is devoid of the 'ActivityEngagement' column. The inclusion of PCA components in this dataset allows us to explore patterns in special activities while maintaining a focus on numerical features.

In the concluding stages of our preprocessing journey, we embark on a critical step—Segmentation. Recognizing the diverse nature of our dataset, we strategically divide our features into two distinct perspectives: socio-demographic and engagement. Socio-demographic features, encompassing Age, Income, and Gender, provide essential insights into customer characteristics, while engagement features, including NumberOfFrequencies, AttendedClasses, RealNumberOfVisits, and others, shed light on consumption behavior and attendance patterns. This segmentation enhances interpretability and simplifies subsequent clustering analyses. After dividing the dataset into these perspectives we selectively retain key demographic information. This focused approach ensures that clustering efforts align with our objectives, facilitating a more nuanced understanding of customer segments based on socio-demographic and engagement attributes. These steps lay the groundwork for insightful clustering analyses, setting the stage for uncovering meaningful patterns within our data.

3.Clustering

In this clustering phase, we began by defining utility functions crucial for evaluating different clustering solutions. The functions include calculating the sum of squared distances and determining the R2 score, providing valuable metrics to assess the quality of clustering. Moreover, we implemented a function to compute the inertia and silhouette scores for a range of cluster numbers, aiding in the identification of an optimal cluster count.

Moving on to hierarchical clustering, we introduced a function to visualize dendrograms with specified linkage methods and distance thresholds, enabling a comprehensive exploration of the data's hierarchical structure.

Subsequently, we delved into clustering the socio-demographic features of our dataset, focusing on key variables such as age, income, and gender. To facilitate this, we utilized the K-Prototypes algorithm, capable of handling mixed numeric and categorical features. The elbow method was employed to determine an optimal number of clusters, and we chose five clusters for subsequent analysis.

Upon conducting the K-Prototypes clustering, we computed silhouette scores to evaluate the quality and coherence of the obtained clusters. The silhouette score of 0.116142 indicates a moderate level of separation and cohesion among the clusters, demonstrating the effectiveness of the chosen approach in segmenting the socio-demographic features. These clusters will serve as a foundation for further analysis and targeted marketing strategies tailored to distinct customer segments.

In the Socio-Demographic segment, we applied multiple clustering algorithms to identify distinct customer groups based on features such as age, income, and gender. First, we employed hierarchical agglomerative clustering using Gower distance to calculate the dissimilarity matrix. The dendrogram visualization aided in determining an optimal number of clusters, and we chose five clusters for further analysis. Silhouette scores were computed for K-Prototypes, Hierarchical Clustering (HC), and K-Medoids to evaluate their performance on the Socio-Demographic dataset.

Notably, the HC algorithm exhibited a high silhouette score of 0.602889, suggesting clear separation and cohesion within the clusters. K-Medoids also yielded a strong silhouette score of 0.614813, indicating well-defined clusters. Both methods outperformed K-Prototypes in this specific context.

The chosen number of clusters, five, was assigned to the original dataset, providing cluster labels for each data point. These labels, along with the Socio-Demographic features, allow for targeted analysis and tailored marketing strategies tailored to each customer segment.

Moving to the Engagement Features segment, we applied K-Means clustering due to its suitability for purely numeric data. The Elbow Method and Silhouette scores were employed to determine the optimal number of clusters. Silhouette scores indicated that two clusters were preferable for Engagement features, showcasing a higher silhouette score compared to other cluster counts.

The K-Means algorithm was then applied to assign cluster labels to the Engagement dataset. Principal Component Analysis (PCA) was subsequently utilized to reduce the dimensionality of the data and facilitate visualization in a scatter plot. The plot visually represents the clusters in a lower-dimensional space, aiding in the interpretation and analysis of customer engagement patterns.

In conclusion, these clustering approaches have provided valuable insights into customer segmentation based on Socio-Demographic and Engagement features. The identified clusters can serve as a foundation for targeted marketing strategies, allowing businesses to tailor their approach to the specific needs and characteristics of distinct customer groups.

4.Cluster Profiling and analysis

Cluster profiling is a crucial step in understanding the patterns and characteristics within different segments of the customer base and offers a comprehensive approach to cluster profiling, focusing on both demographic and engagement perspectives. The analysis is conducted using K-means clustering, which divides the customer base into distinct groups based on specific features.

Starting with demographic clustering, the code reveals five clusters, each representing a unique segment of the population. However, the analysis indicates that these clusters do not exhibit a clear distinction related to genders. Instead, certain activities such as FitnessActivities and WaterActivities contribute significantly to the clustering, suggesting that these activities strongly influence the customer segmentation.

The exploration of income and age within clusters unveils a specific segment (Cluster 1) characterized by lower valuation, likely representing a younger and lower-income population. This group may require targeted marketing strategies and special offerings to reduce dropout rates. Additionally, variations in attendance features like TimeSinceLastVisit and DaysWithoutFrequency point to specific behavioral patterns in clusters 0 and 3, emphasizing the need for tailored promotions to encourage consistent participation.

Moving on to the engagement perspective, K-means clustering reveals two clusters, indicating distinct levels of customer loyalty and health consciousness. The analysis illustrates that one group tends to attend the fitness center less frequently, while the other demonstrates a higher level of loyalty and engagement. This insight is valuable for devising marketing strategies and promotions to boost engagement among different customer segments.

The t-SNE visualizations provide a two-dimensional representation of clusters, allowing for a more intuitive understanding of their distribution. The Decision Tree models further contribute to the analysis by showcasing the key features influencing cluster assignments. The feature importance analysis highlights the factors contributing to the prediction accuracy of the Decision Tree models.

In conclusion, the cluster profiling and analysis offer meaningful insights into the customer base, enabling the identification of distinct segments with unique characteristics. These insights can inform targeted marketing strategies, customer retention efforts, and the development of personalized offerings. Additionally, the combination of t-SNE visualizations and Decision Tree models enhances the interpretability of the clustering results. The decision tree here visualized is an excellent tool for interpretability of our business case, being a valid way of predicting possible dropout customers with the surveillance of their features. We can see that some of the deciding factors for customer loyalty, besides Gender, are TimeSinceLastVisit, NumberOfFrequencies and Age. All these features should be taken into account for marketing approaches for customer retention.

Overall, this comprehensive approach provides a solid foundation for data-driven decision-making in marketing and customer relationship management.

5. References

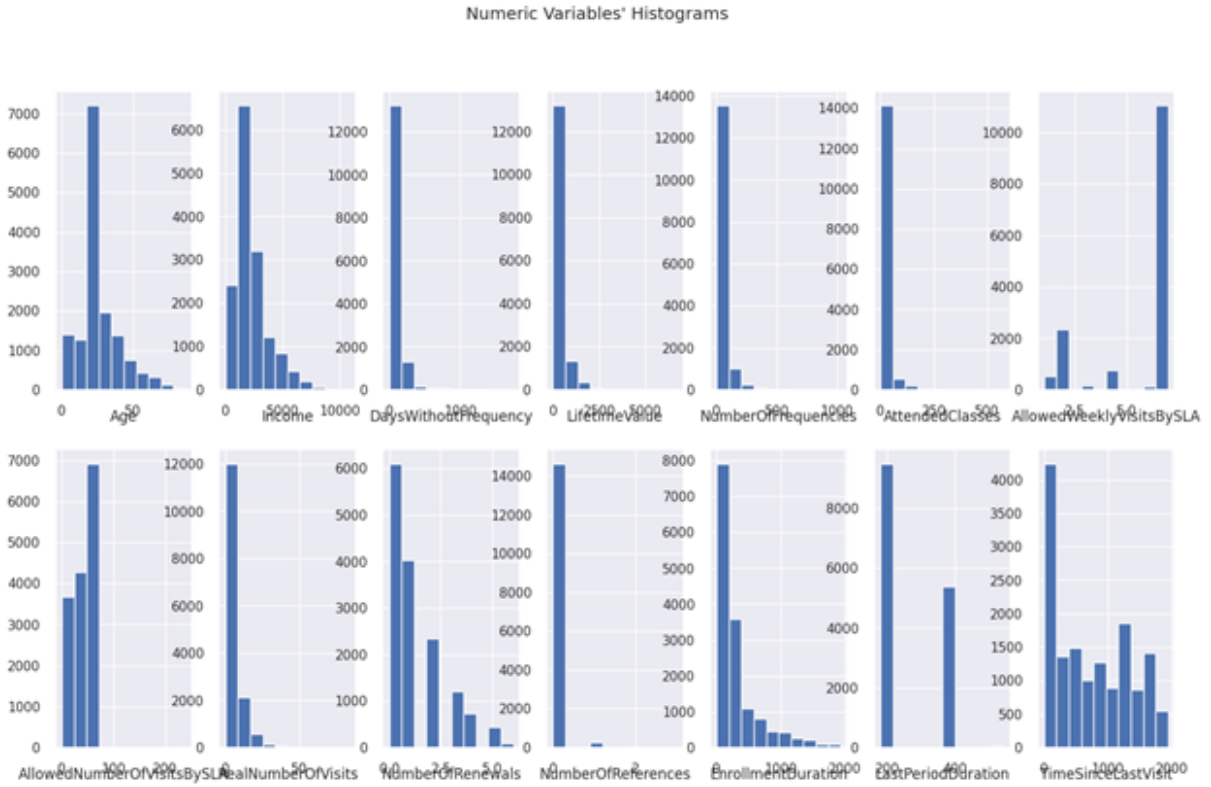
Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

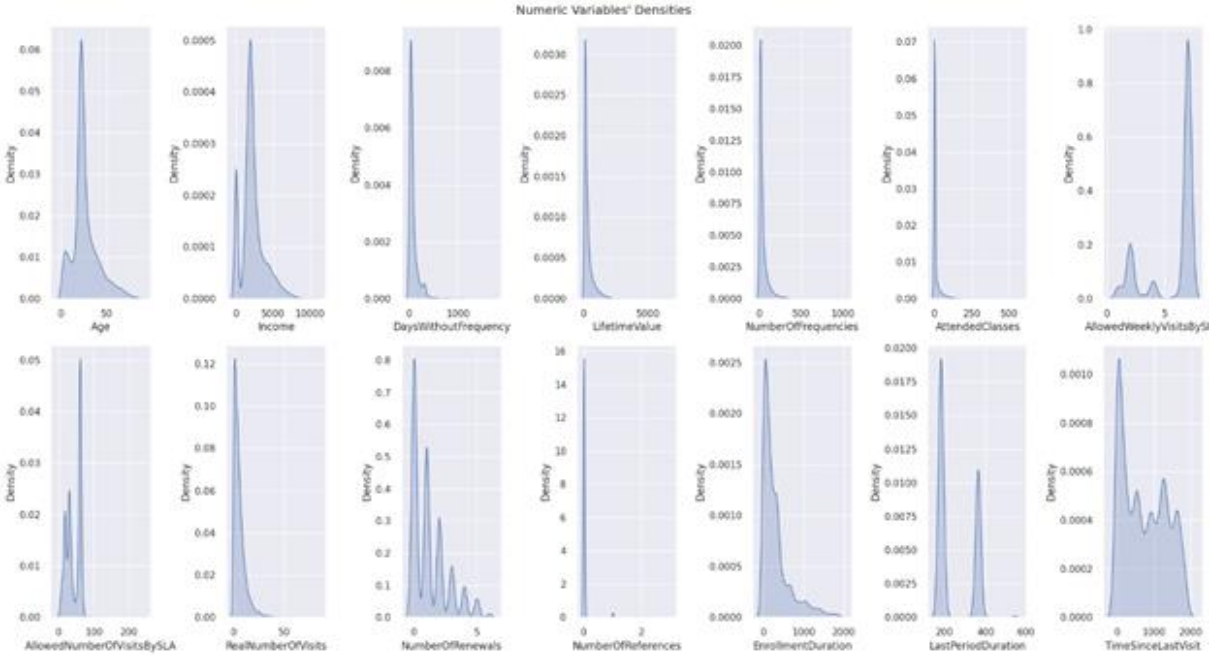
Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

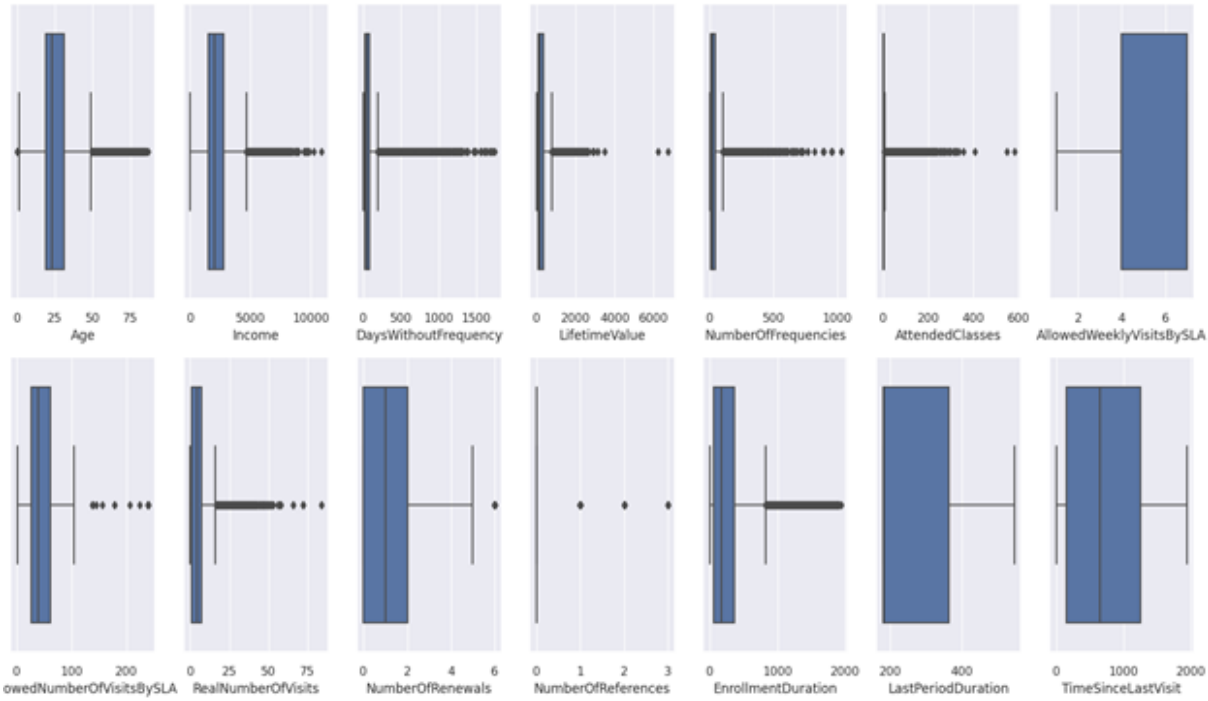
6. Appendix

Exploratory Data Analysis

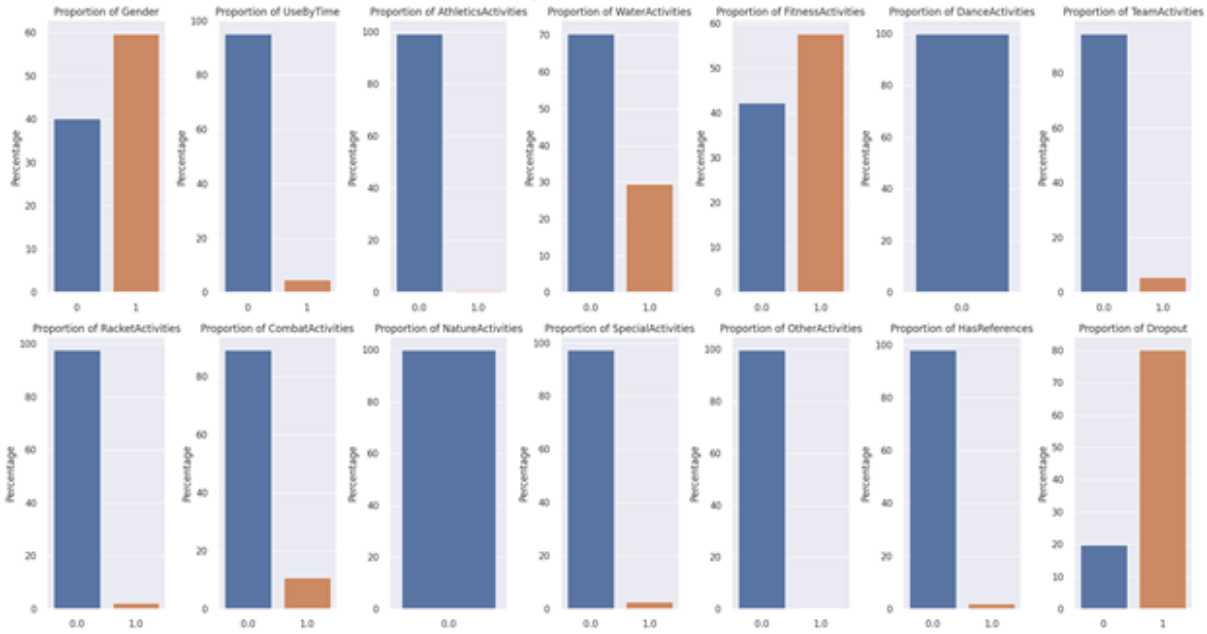


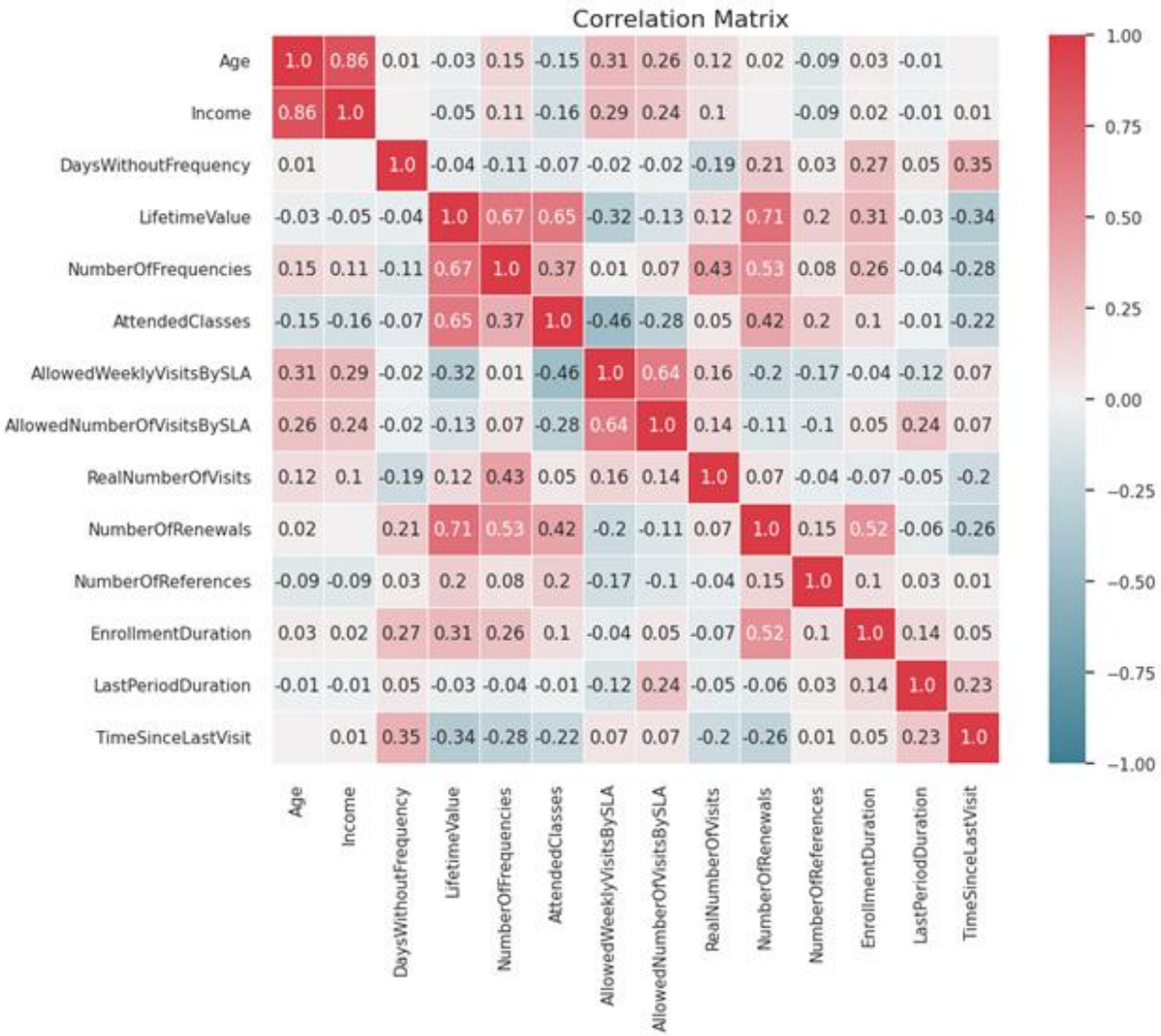


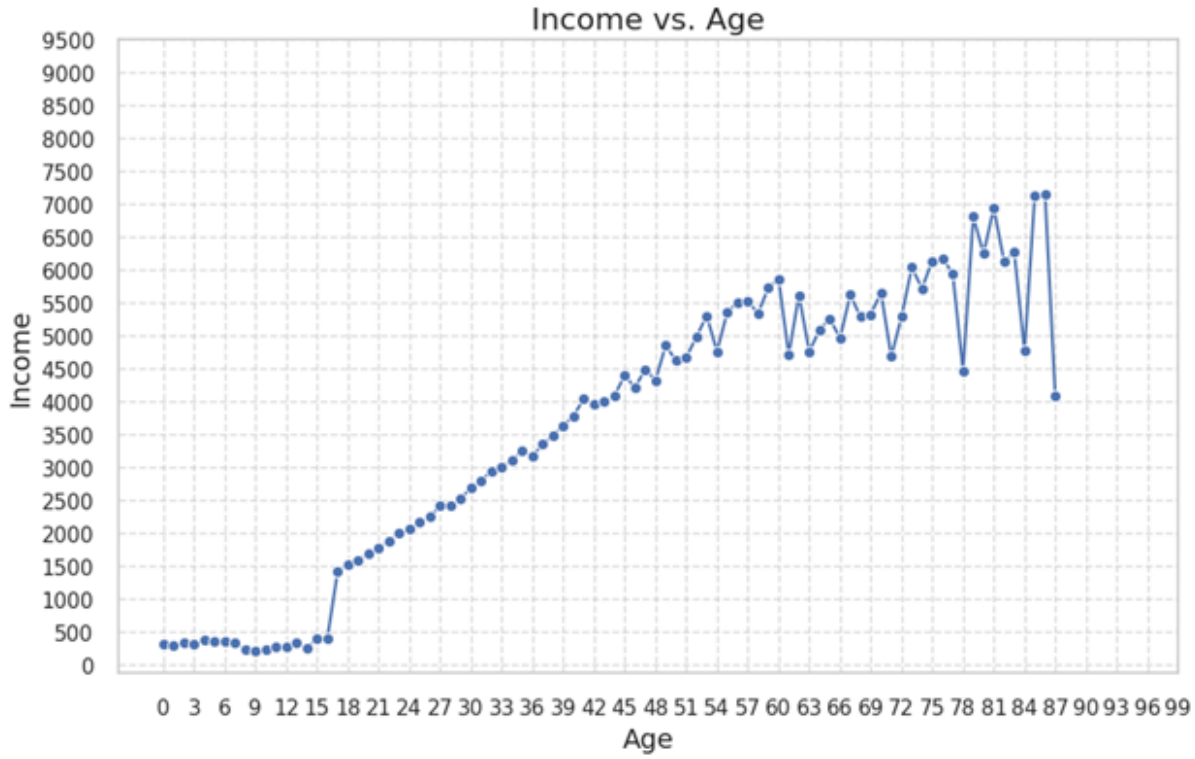
Numeric Variables' Box Plots



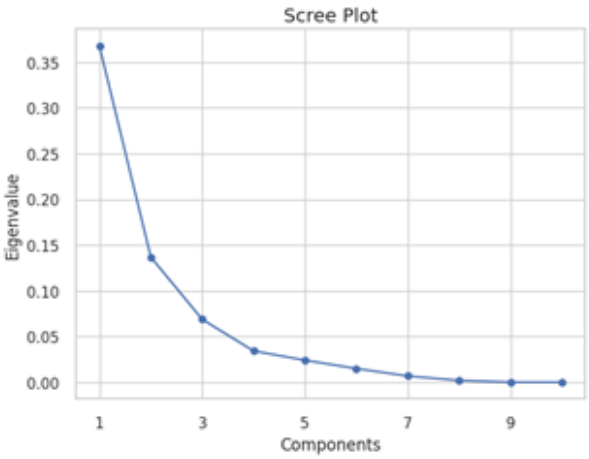
Categorical Variables' Bar Plots

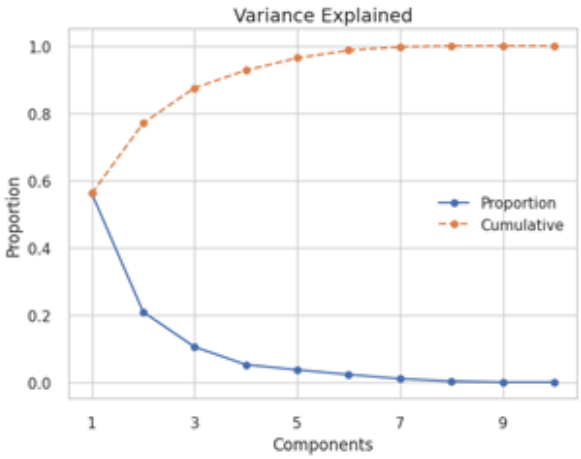



Correlation Matrix
Age 1.0 0.86 0.01 -0.03 0.15 -0.15 0.31 0.26 0.12 0.02 -0.09 0.03 -0.01
Income 0.86 1.0 -0.05 0.11 -0.16 0.29 0.24 0.1 -0.09 0.02 -0.01 0.01
DaysWithoutFrequency 0.01 1.0 -0.04 -0.11 -0.07 -0.02 -0.02 -0.19 0.21 0.03 0.27 0.05 0.35
LifetimeValue -0.03 -0.05 -0.04 1.0 0.67 0.65 -0.32 -0.13 0.12 0.71 0.2 0.31 -0.03 -0.34
NumberOfFrequencies 0.15 0.11 -0.11 0.67 1.0 0.37 0.01 0.07 0.43 0.53 0.08 0.26 -0.04 -0.28
AttendedClasses -0.15 -0.16 -0.07 0.65 0.37 1.0 -0.46 -0.28 0.05 0.42 0.2 0.1 -0.01 -0.22
AllowedWeeklyVisitsBySLA 0.31 0.29 -0.02 -0.32 0.01 -0.46 1.0 0.64 0.16 -0.2 -0.17 -0.04 -0.12 0.07
AllowedNumberOfVisitsBySLA 0.26 0.24 -0.02 -0.13 0.07 -0.28 0.64 1.0 0.14 -0.11 -0.1 0.05 0.24 0.07
RealNumberOfVisits 0.12 0.1 -0.19 0.12 0.43 0.05 0.16 0.14 1.0 0.07 -0.04 -0.07 -0.05 -0.2
NumberOfRenewals 0.02 0.21 0.71 0.53 0.42 -0.2 -0.11 0.07 1.0 0.15 0.52 -0.06 -0.26
NumberOfReferences -0.09 -0.09 0.03 0.2 0.08 0.2 -0.17 -0.1 -0.04 0.15 1.0 0.1 0.03 0.01
EnrollmentDuration 0.03 0.02 0.27 0.31 0.26 0.1 -0.04 0.05 -0.07 0.52 0.1 1.0 0.14 0.05
LastPeriodDuration -0.01 -0.01 0.05 -0.03 -0.04 -0.01 -0.12 0.24 -0.05 -0.06 0.03 0.14 1.0 0.23
TimeSinceLastVisit 0.01 0.35 -0.34 -0.28 -0.22 0.07 0.07 -0.2 -0.26 0.01 0.05 0.23 1.0
Age
Income
DaysWithoutFrequency
LifetimeValue
NumberOfFrequencies
AttendedClasses
AllowedWeeklyVisitsBySLA
AllowedNumberOfVisitsBySLA
RealNumberOfVisits
NumberOfRenewals
NumberOfReferences
EnrollmentDuration
LastPeriodDuration
TimeSinceLastVisit
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



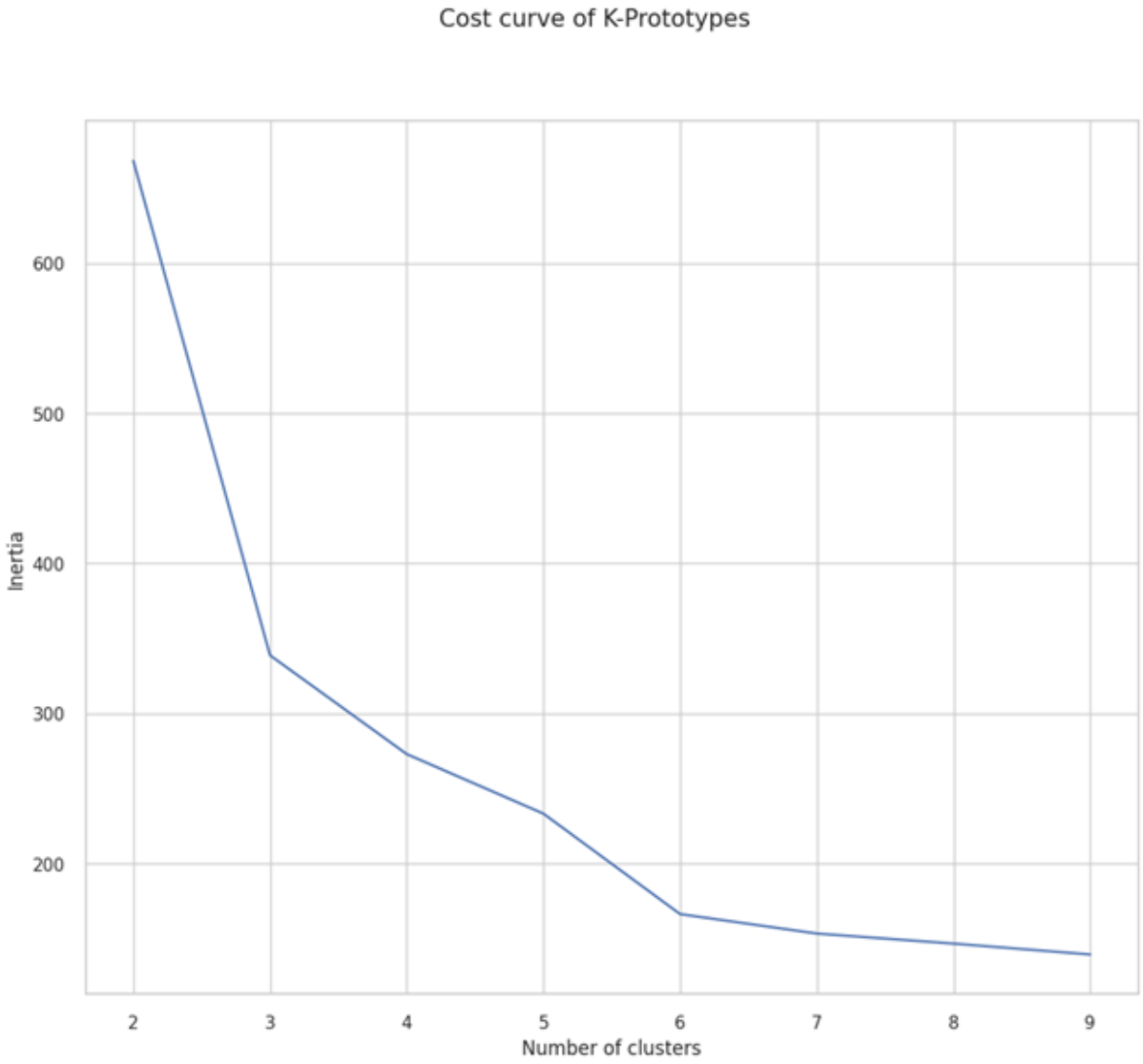
Binary PCA Reduction (Activities)

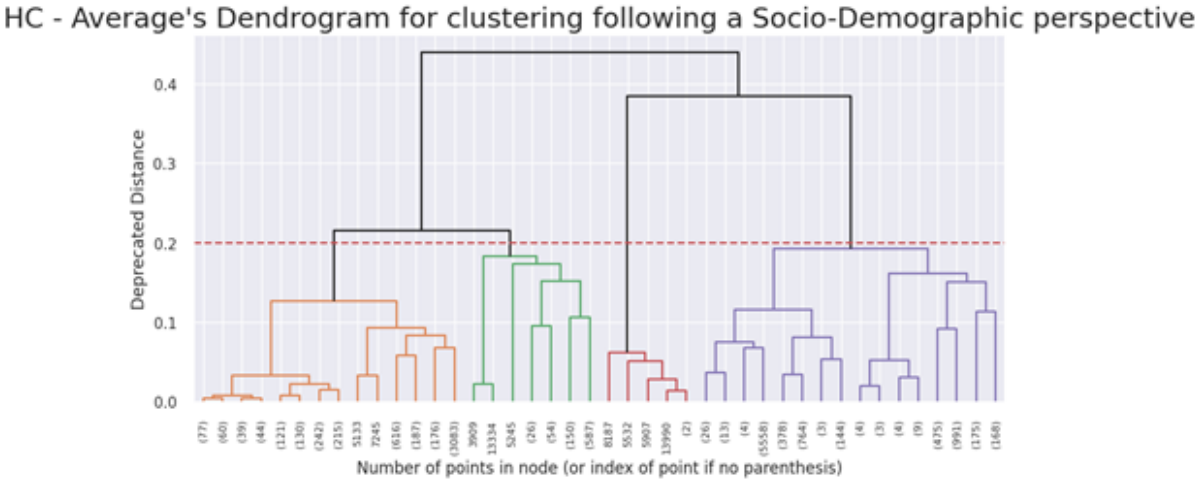




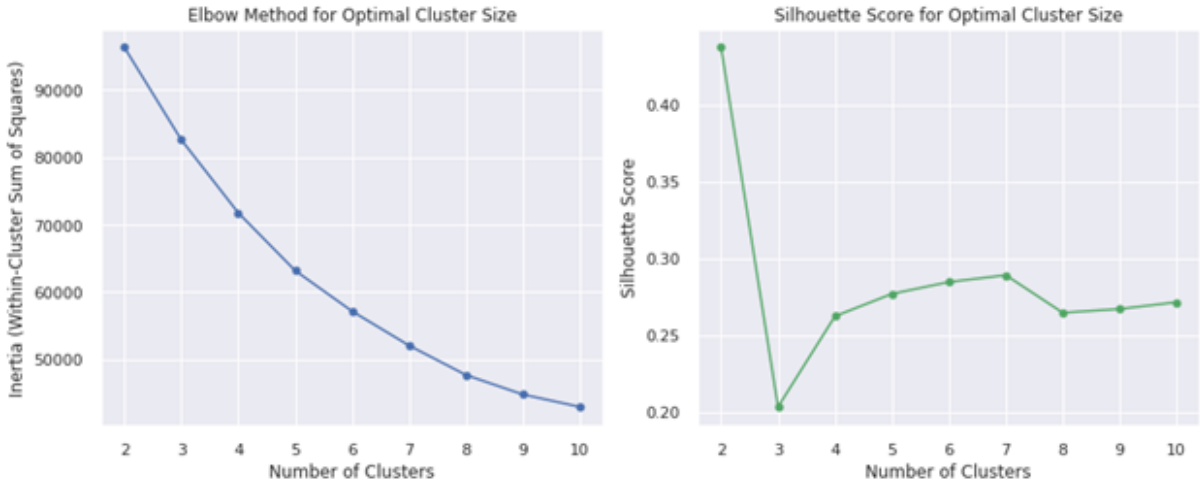
	PCA_Activity_0	PCA_Activity_1	PCA_Activity_2
DanceActivities	nan	nan	nan
OtherActivities	0.015190	0.032980	0.046786
NatureActivities	nan	nan	nan
RacketActivities	0.052113	0.132663	0.207886
AthleticsActivities	0.028628	0.051197	0.068511
WaterActivities	0.862394	-0.448882	-0.187930
FitnessActivities	-0.930612	-0.308646	-0.144730
TeamActivities	0.130085	0.252016	0.790135
CombatActivities	0.103012	0.841113	-0.487318
SpecialActivities	0.011738	0.051514	0.113736

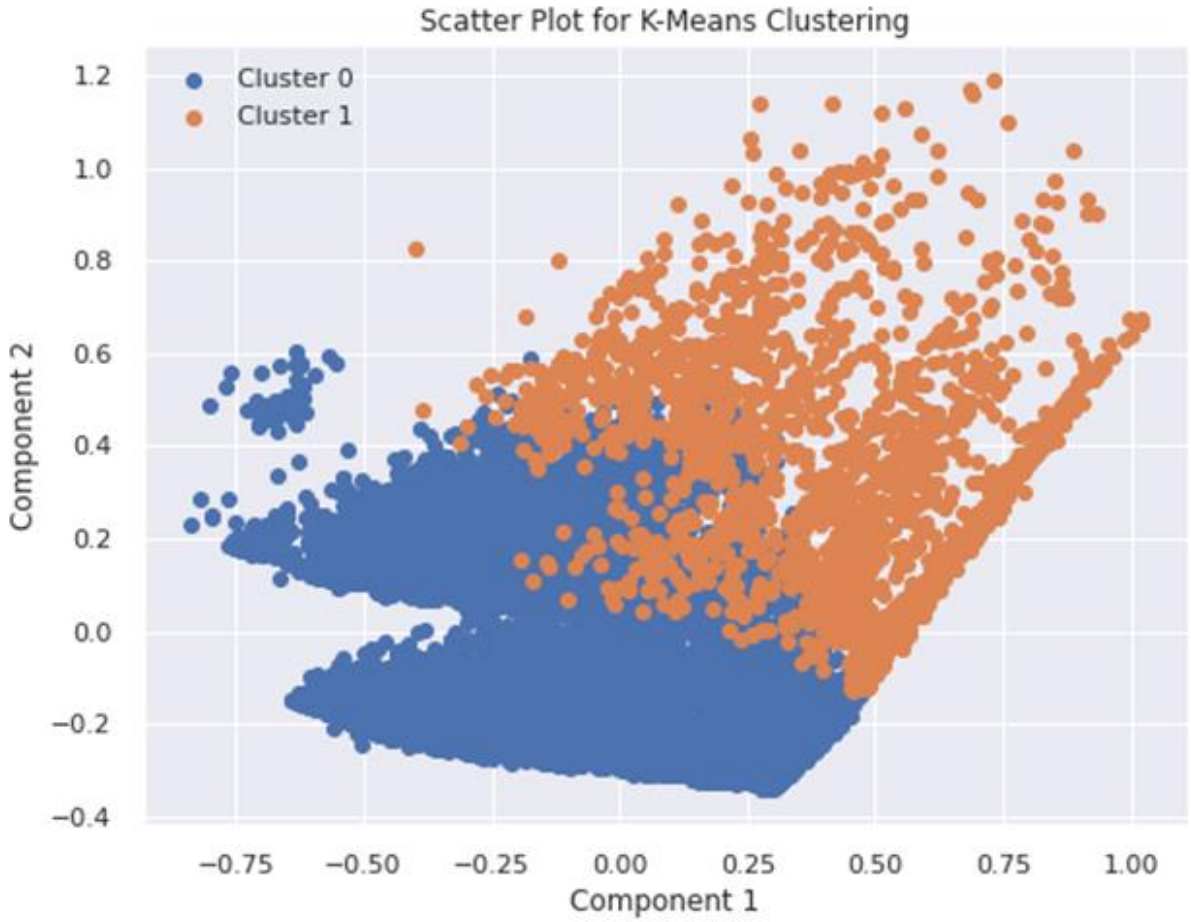
K-prototype clustering test for demographics perspective



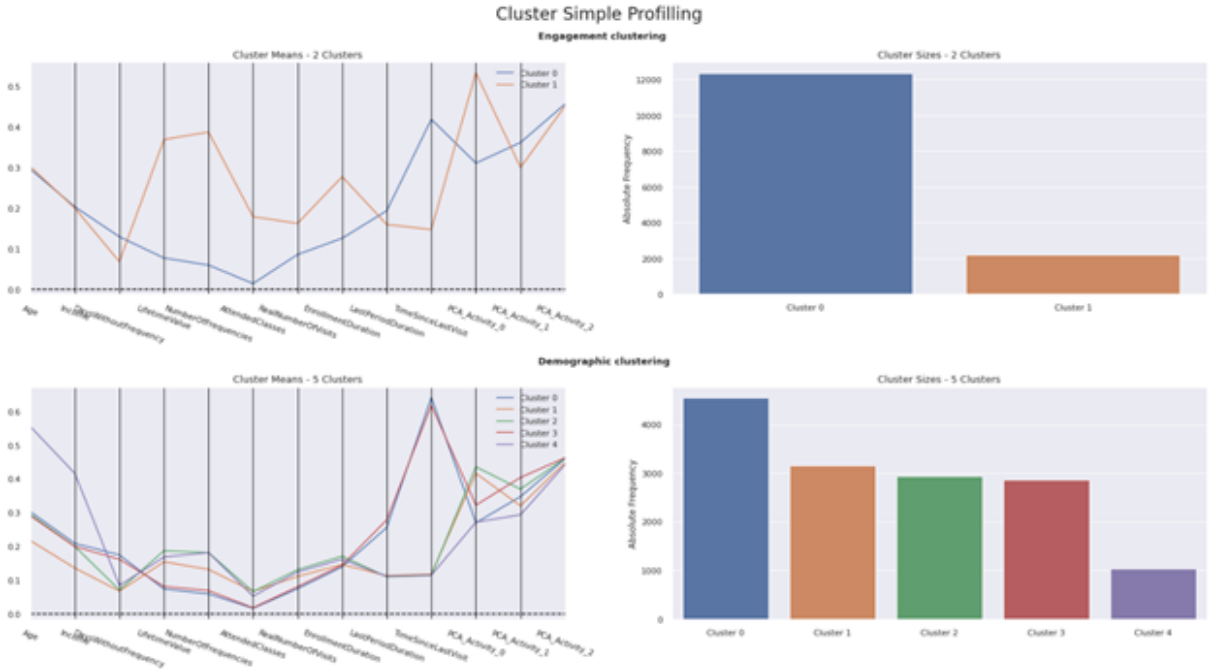


K-means clustering test for engagement perspective

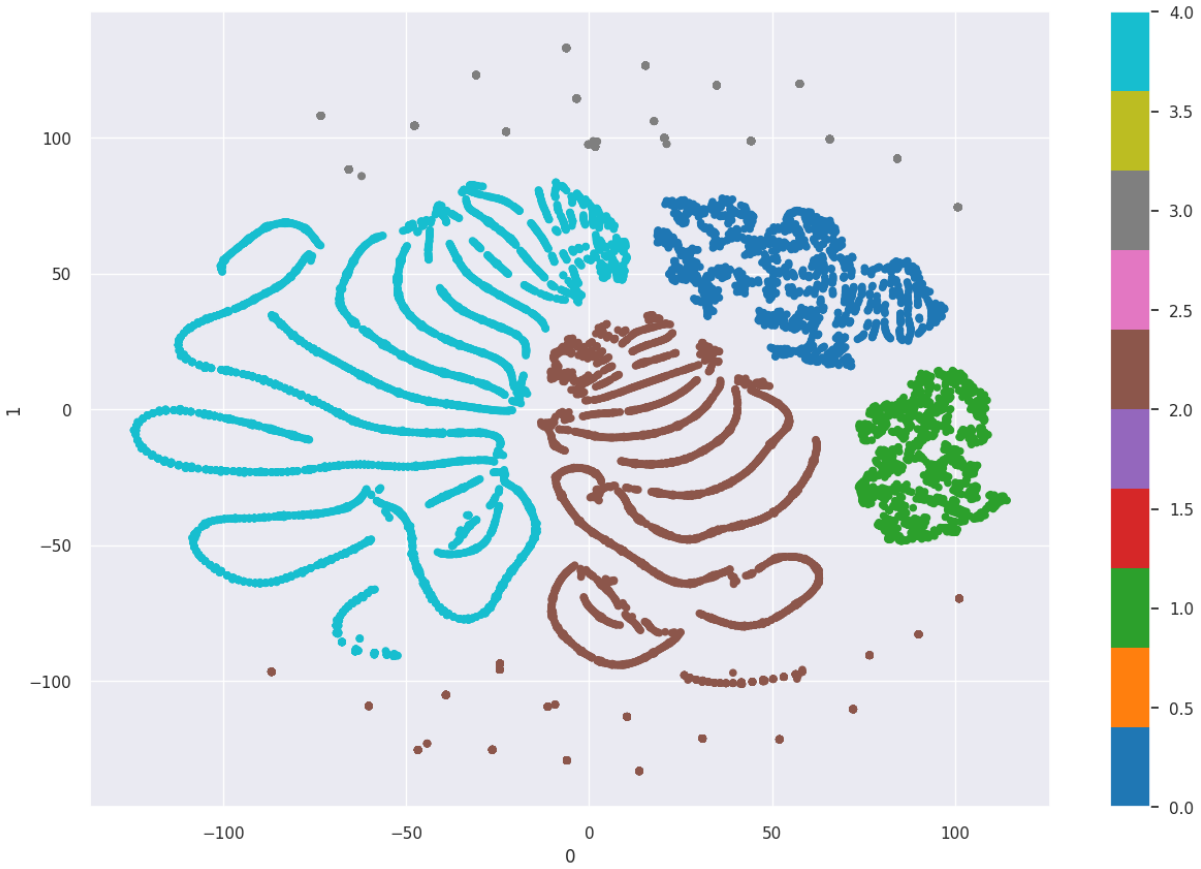




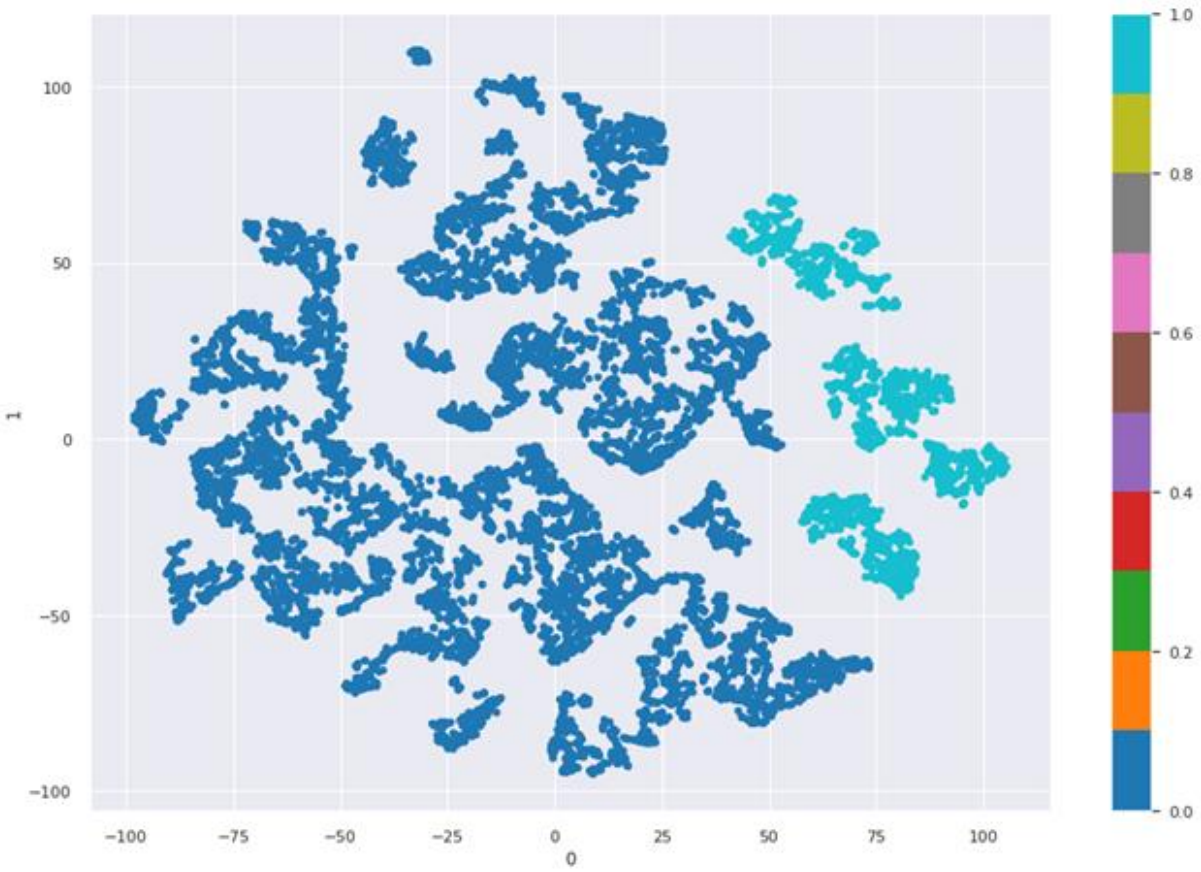
Profiling both perspectives



T-SNE plot for demographics perspective



T-SNE plot for engagement perspective



Decision Tree plot