CS 6220: Final Project

HDM Data Analysis

1st Hao Sheng Ning
Northeastern University
Portland, Maine, USA
ning.ha@northeastern.edu

2nd Raja Shiva Ram Patel
Northeastern University
Portland, Maine, USA
patel.rajas@northeastern.edu

***Abstract*—In this report, we take a closer look at the Home Delivery Meals (HDM) program managed by the Southern Maine Agency on Aging (SMAA) using a multidimensional data analysis approach. Our goal is to unravel the complex connections between client financial situations, social dynamics, and health factors. The study aims to reveal subtle patterns within the dataset, offering practical insights for the SMAA. By digging into the data, we aim to identify the demographics that need assistance the most. Armed with this knowledge, the SMAA can make informed decisions on resource allocation, ensuring that the HDM program makes a significant impact on the well-being of the aging population in Southern Maine. Our methodology involves Exploratory Data Analysis (EDA) for initial pattern discovery, Support Vector Machine (SVM) and Decision Tree models to uncover intricate relationships, Principal Component Analysis (PCA) for simplifying complex data, and K-means Clustering to identify distinct client segments. Through this blend of analytical approaches, we aspire to gain a comprehensive understanding of the various factors influencing the utilization of the HDM program.**

I. Introduction

The landscape of aging populations presents a growing challenge for social service organizations, requiring innovative approaches to address the diverse and evolving needs of seniors. Among the critical services offered to the elderly, the Home Delivery Meals (HDM) program plays a pivotal role in ensuring nutrition, well-being, and independence. Recognizing the significance of efficient resource allocation to maximize the impact of such programs, this report presents a comprehensive data analysis conducted in collaboration with the Southern Maine Agency on Aging (SMAA).

As demographics continue to shift, it becomes imperative for organizations like SMAA to adapt and tailor their services to the specific requirements of their constituents. The HDM program, designed to provide nutritious meals to older adults, serves as a lifeline for many, particularly those facing financial, social, or health challenges. This study aims to unravel the intricate relationships between client demographics and the utilization of HDM services, with a specific focus on informing the SMAA about the demographics most in need.

In the pursuit of this objective, our research employs a multifaceted approach, integrating Exploratory Data Analysis (EDA), Support Vector Machine (SVM), Decision Tree, Principal Component Analysis (PCA), and K-means Clustering methodologies. These advanced analytical tools will enable us to delve deep into the data, extracting valuable insights that go beyond surface-level observations. By analyzing these insights on clients' financial, social, and health backgrounds, this study seeks to uncover patterns and correlations that will guide the Southern Maine Agency on Aging in strategically allocating resources. The overarching goal is to enhance the effectiveness of the HDM program, ensuring that those who are most in need receive the support required to maintain their well-being and independence.

As we embark on this data-driven journey, the findings and recommendations derived from our analysis are poised to not only refine resource allocation within the HDM program but also contribute to the broader mission of the SMAA. Ultimately, this research aims to empower the Southern Maine community in better serving its aging population, fostering a more inclusive, responsive, and targeted approach to support the diverse needs of older adults.

II. Methods/Analysis

The initial phase of our analysis involved a comprehensive exploration of the dataset to uncover patterns, trends, and potential outliers. This Exploratory Data Analysis (EDA) serves as the foundation for deeper insights into the dynamics of the Home Delivery Meals (HDM) program.

We started the Exploratory Data Analysis by gaining a holistic understanding of the dataset's structure, dimensions, and basic statistics. The dataset, sourced from the Southern Maine Agency on Aging (SMAA), comprises a diverse set of variables capturing information about client demographics, financial status, social dynamics, health conditions, and program utilization.

The demographic statistics are as follows:

1. Distribution of age
2. Gender distribution
3. Race information

1. Distribution of age: Fig. 1 is a histogram that shows the distribution of age. The x-axis represents age, and the y-axis represents frequency. The tallest bar is in the 40-59 age range.

2. Gender Distribution: Fig. 2 is a pie chart that shows the gender distribution of a group of people. The pie chart is divided into two halves, one for male and one for female.

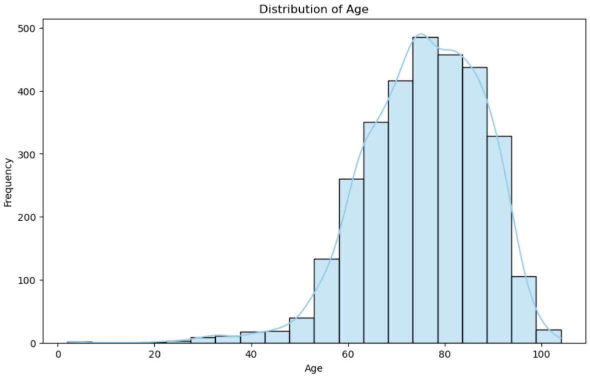


Fig. 1: Age Distribution

The pie chart shows that 57.3% of the people are female and 42.7% of the people are male.

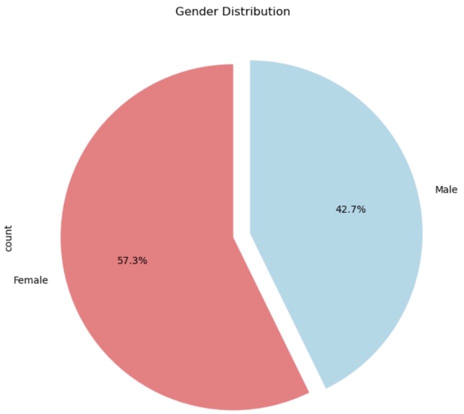
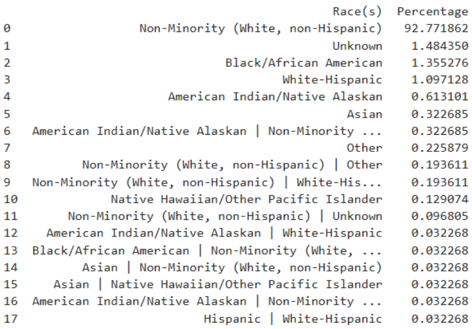


Fig. 2: Gender Distribution

3. Race information: Fig. 3 is the visualization of the racial distribution of respondents indicating that the majority, at 92.77%, identify as Non-Minority (White, non-Hispanic). Other notable groups include those who identified as Black/African American (1.36%) and White-Hispanic (1.10%). The category ”Unknown” represents 1.48% of respondents. Various other racial categories, each comprising less than 1%, contribute to the overall diversity of the sample.



```
                                            Race(s)  Percentage
0              Non-Minority (White, non-Hispanic)   92.771862
1                                          Unknown    1.484350
2                           Black/African American    1.355276
3                                   White-Hispanic    1.097128
4                   American Indian/Native Alaskan    0.613101
5                                            Asian    0.322685
6   American Indian/Native Alaskan | Non-Minority ...    0.322685
7                                            Other    0.225879
8       Non-Minority (White, non-Hispanic) | Other    0.193611
9   Non-Minority (White, non-Hispanic) | White-His...    0.193611
10            Native Hawaiian/Other Pacific Islander    0.129074
11    Non-Minority (White, non-Hispanic) | Unknown    0.096805
12   American Indian/Native Alaskan | White-Hispanic    0.032268
13  Black/African American | Non-Minority (White, ...    0.032268
14      Asian | Non-Minority (White, non-Hispanic)    0.032268
15     Asian | Native Hawaiian/Other Pacific Islander    0.032268
16  American Indian/Native Alaskan | Non-Minority ...    0.032268
17                      Hispanic | White-Hispanic    0.032268
```

Fig. 3: Race Information

A. Data Format

Data is read from csv file into ipynb as Pandas Dataframe object. Its values are converted to list or numpy at times to satisfy algorithm it will go through.

B. Data Cleaning

As observed from EDA, there are definitely some outliers, most likely errors in the dataset, for example someone weights 13540 lbs and someone who is 0.66 inches tall. All the errors seem to be found in numberic data, such as weights and height, thus we remove data that sastifies following criteria:

- Weight greater than 1000 lbs
- Height less than 24 inches
- Height greater than 110 inches

There are also many rows with empty values. For simplicity, we removed them. That amounts to 2731 out of 3099 total entries. That means we are left only with 11.2% (368) of total data to do remainder of the test, which is still sufficient.

C. Exploratory Data Analysis

Distribution of Title III eligibility for males and females: Fig. 4 and Fig. 5 are pie charts that show the distribution of Title III eligibility for males and females respectively. The pie chart is divided into six slices, each representing a different eligibility category.

For males, the largest slice (85.9%) represents those who are 60+ years old and in the greatest socio-economic need. The next largest slice (9.0%) represents those who are not eligible for Title III services. The remaining slices represent those who are 60+ years old and clients of Adult Protective Services (0.8%), spouses (any age) of current Home Delivered Meals participants (3.1%), and under 60 years old, disabled, and living with a current Home Delivered Meals participant (1.2%).

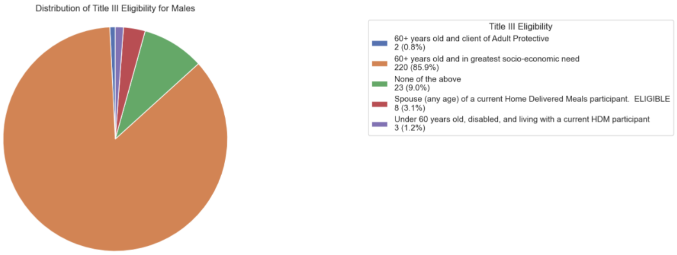


Fig. 4: Distribution Of Title III Eligibility For Males

For females, the largest slice (90.1%) also represents those who are 60+ years old and in the greatest socio-economic need. The next largest slice (7.9%) represents those who are not eligible for Title III services. The remaining slices represent those who are 60+ years old and clients of Adult Protective Services (0.3%), spouses (any age) of current Home Delivered Meals participants (1.2%), and under 60 years old, disabled, and living with a current Home Delivered Meals participant (0.6%).

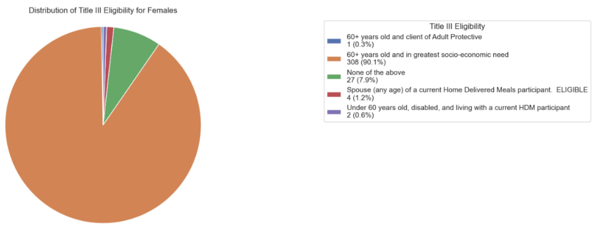


Fig. 5: Distribution Of Title III Eligibility For Females

Overall, the pie chart shows that the majority of both males and females who are eligible for Title III services are 60+ years old and in the greatest socio-economic need.

Financial and Housing Concerns: Fig. 6 is a stacked count-plot that shows the number of people who have selected each expense as their most concerning expense, grouped by their ability to afford bills. The x-axis shows the ability to afford bills, and the y-axis shows the count of people in each category. The bars are stacked to show the total number of people who have selected each expense as their most concerning expense, regardless of their ability to afford bills. This visual shows the breakdown of most concerning expenses based on the ability to afford bills. It reveals that food, rent/utility bills, and medical expenses are the top concerns for those struggling financially.

Key Findings:

- Food is the most concerning expense for people who are unable to afford bills.
- Rent/utility bills and medical expenses are also major concerns for people who are unable to afford bills.
- People who are able to afford bills are more likely to be concerned about other expenses, such as transportation and education.

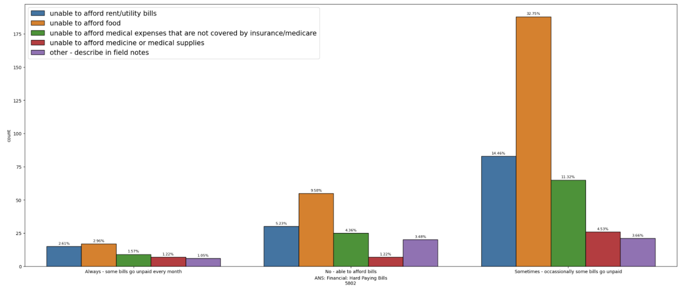


Fig. 6: Distribution of Most Concerning Expenses by Ability to Afford Bills

Correlation Heat-map: Fig. 7 is a correlation heat-map depicting the relationships between various numerical variables in your data. It utilizes color intensity and numerical values to represent the strength and direction of these relationships. The correlation coefficient is a measure of the linear relationship between two variables. It can range from -1 to 1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation.
The x-axis and y-axis represent the same set of variables, allowing for easy comparison across the matrix. Each cell within the heat map corresponds to the correlation coefficient between the two variables represented by its row and column. Correlation coefficients are displayed within each cell, rounded to two decimal places, providing a quantitative measure of the relationship.

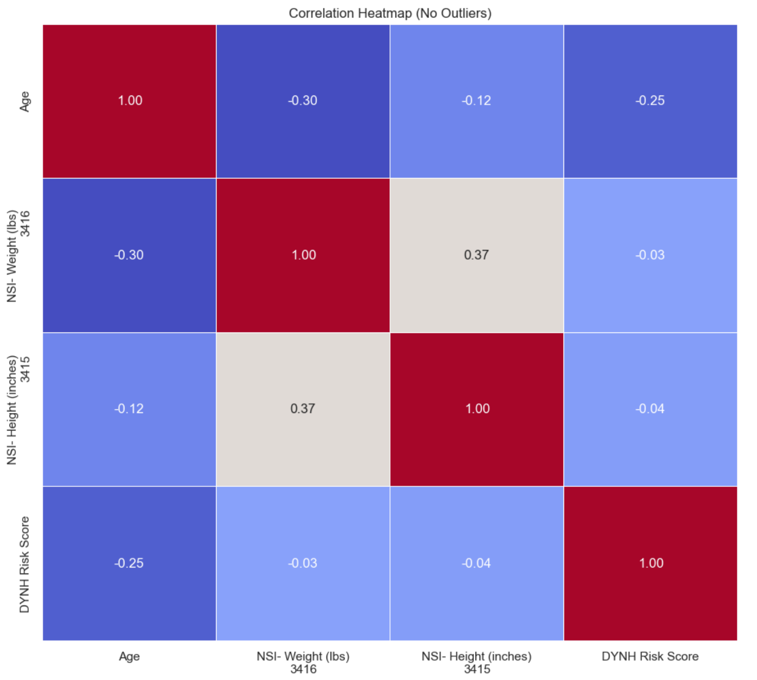


Fig. 7: Correlation Heat Map

Key Observations:

- There is a positive correlation between NSI- Weight (lbs) and NSI- Height (inches). This is reflected by the grey color and a correlation coefficient (approximately 0.37) in the corresponding cell.
- Other correlations appear much weaker, with most cells displaying lighter shades of blue, closer to zero.
- This suggests that NSI- Weight (lbs) and NSI- Height (inches) are the only two variables with a statistically significant relationship in this dataset.

D. Median and T-test

A t-test is a statistical method used to determine if there is a significant difference between the means or medians of two groups. It is particularly relevant when comparing middle points from two independent samples, such as comparing the average scores of two different groups or assessing the impact of a treatment on a group compared to a control. The t-test evaluates whether the observed difference between the group means is statistically significant or if it could have occurred by random chance. The p-value associated with it quantifies this probability, specifically the likelihood of obtaining the observed difference or a more extreme one under the assumption that there is no real difference between the groups. A lower p-value, typically below the conventional threshold of 0.05, indicates greater evidence against the null hypothesis of no difference, suggesting statistical significance. There are variations of the t-test, including the independent samples t-test for comparing means of two separate groups and the paired samples t-test for comparing means within the same group under different conditions. By analyzing the t-statistic and degrees of freedom, researchers can ascertain whether the observed differences are likely due to a real effect or if they could be the result of random variability in the data. Overall, the t-test is a valuable tool for hypothesis testing and concluding differences between groups in various scientific and research contexts.

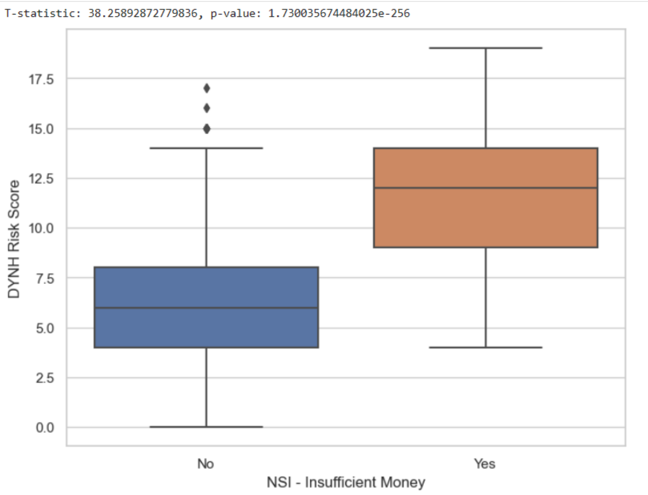
Fig. 8-11 depicts the variability in "DYNH Risk Score" across distinct groups characterized by different variable values. The significance of the observed differences is gauged by the p-value displayed atop each figure, reflecting the probability that the differences occurred by pure chance. As mentioned above, a p-value below 0.05 is often indicative of statistical significance in the observed variations, meaning the different variable values did indeed have an impact on the target result. All 4 t-tests in the figures showed a p-value smaller than the threshold, the smallest one being the InsufficientMoney variable, which shows financial strength is a potent factor in determining one's risk score. Notably, Figure 8 stands out as it highlights a substantial difference in "DYNH Risk Score" among clients from different towns, supported by a notably low p-value. However, interpreting this difference requires additional information or consultation with experts, as it cannot be fully explained without further context.

Given that there are numerous variables beyond those depicted in Figures 8-11, we have developed the following web interface: https://hsning.github.io/Projects/HDMDemo.html This platform enables users to select any variable of interest, allowing them to explore different midpoints and delve into t-tests between various values.

III. Results

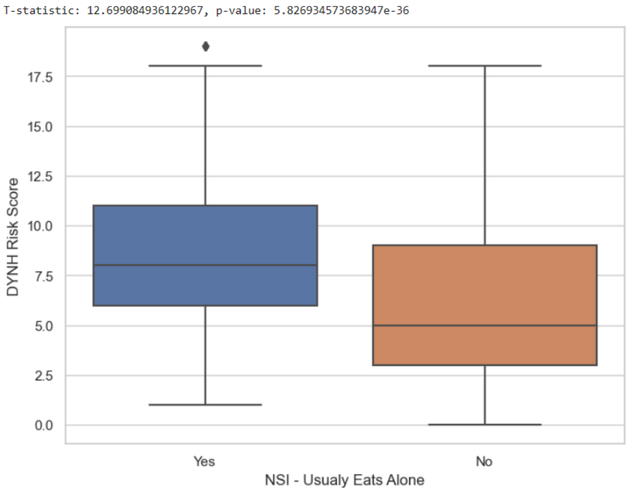
As the raw result is very big, we find it more suitable to just include the tables/graphs in the report and leave the raw result in the embedded ipynb file.

As the aim of SMAA is to reduce the risk score for their clients. It is important to know which factors have the strongest effect on achieving this goal, so they can have an area of focus in terms of resource allocation. We developed two methods to demonstrate feature importance: 1) SVM 2) Decision Tree. Both models use input variables to predict the target variable "DYNH Risk Score".



```
T-statistic: 38.25892872779836, p-value: 1.730035674484025e-256
```

Fig. 8: Insufficient Money

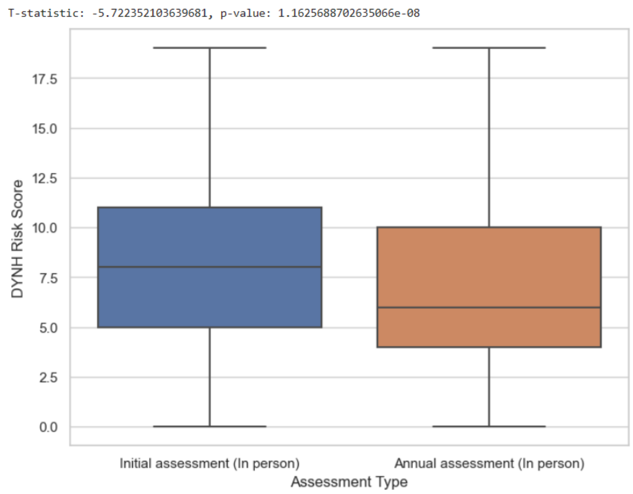


```
T-statistic: 12.699084936122967, p-value: 5.826934573683947e-36
```

Fig. 9: Usually Eats Alone

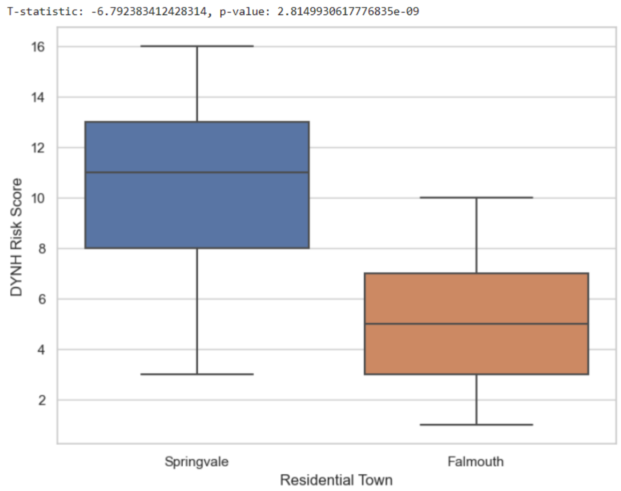
A. SVM

Support Vector Machines (SVM) are primarily used for classification and regression tasks, and they inherently do not provide direct measures of feature importance like some other algorithms, such as decision trees or random forests. However, there are some ways in which SVM-related techniques can indirectly contribute to determining feature importance. One way we are going to achieve this is through analyzing SVM's hyperplane and its coefficients. In linear SVM, the decision boundary is represented by a hyperplane. The coefficients of the hyperplane equation indicate the contribution of each feature to the decision. Larger absolute values of the coefficients suggest more importance. The SVM classification results in an accuracy of **0.810**. Table I shows some features with the highest importance after SVM is applied to the dataset:



```
T-statistic: -5.722352103639681, p-value: 1.1625688702635066e-08
```

Fig. 10: Assessment Types



```
T-statistic: -6.792383412428314, p-value: 2.8149930617776835e-09
```

Fig. 11: Residential Towns

B. Decision Tree

In a decision tree, at each node, the algorithm decides which feature to split the data on. The decision is made based on a criterion such as Gini impurity (for classification) or mean squared error (for regression). The feature that results in the best separation of classes or reduction in variance is considered more important. Decision trees use metrics like information gain to determine the importance of features. Information gain quantifies the improvement in predictive power obtained by splitting the data based on a particular feature. Features with higher information gain are deemed more important. The Decision Tree classification results in an accuracy of **0.972**. Table II shows some features with the highest importance after Decision Tree analysis is applied to the dataset:

Feature	Feature Coefficients
NSI - Takes 3+ Rx Daily	0.0962
ANS: Isolation: How Satisfied With Ability to Visit with Others	0.0955
NSI - Tooth or Mouth Probs	0.0930
ANS: Home Support: Adaptive Equipment Used	0.0874
NSI - Usualy Eats Alone	0.0787
ANS: Chronic Conditions: Reported Conditions	0.0715
NSI - 3+ Alc Drinks Daily	0.070
ANS: Demo: Served During Wartime	0.064
Residential Town	0.0632
ANSI- Height (inches)	0.062

TABLE I: SVM Feature Importance

Feature	Feature Coefficients
IADL Meal Prep	0.1353
ADL Walking	0.0893
ADL Eating	0.0733
ADL Shopping	0.0688
ADL Dressing	0.0583
ADL Transfers	0.0118
ADL Bathing	0.0067

TABLE II: Decision Tree Feature Importance

PCA 1	PCA 2
ANS: Demo: Household Size	ANS: Memory: Other Known Info re Mem
NSI - Insufficient Money	ANS: Memory: Self Concerns
NSI - Usualy Eats Alone	ANS: Memory: Potential Dementia Pt in Reporting
ANS: Demo: Lives Alone	ADL Dressing
ANS: Isolation: How Satisfied With Ability to Visit with Others	ADL Transfers
ANS: Isolation: How Often Feel Left Out	IADL Rx Manage
ANS: Financial: Hard Paying Bills	NSI - Illness Changed Diet

TABLE III: Top 2 PCAs

C. Principal Component Analysis

PCA ranks the principal components in terms of the amount of variance they explain in the data. The first principal component explains the most variance, the second principal component explains the second most, and so on. In this sense, the original features that contribute more to the variance captured by the top principal components can be considered more important. The loading scores in PCA represent the weights assigned to each original feature in the construction of the principal components. Larger loading scores (in absolute value) indicate stronger contributions of the corresponding features to the principal components. Features with higher loading scores are considered more important in explaining the variability in the data.

In the context of numerical PCA, the typical approach involves standardizing the dataset. However, in this case, as the dataset is predominantly categorical, we opted for encoding the data into binary variables. As a result, standardization was not performed in this scenario.

The following Table III shows the two highest PCAs and their original features and their respective weightings:

It is clear from the PCA table and its component variables that PCA1 is more related to the degree of loneliness the client experiences and PCA2 is more to do with their personal health.

D. Clustering

Up to this point, our analysis has been focused on the ”DYNH Risk Score.” However, it is highly probable that there are additional insights beyond ”Risk.” For this reason, we aim to employ an unsupervised machine learning algorithm, such as clustering, to explore the dataset for hidden trends yet to be discovered.

K-means clustering is an unsupervised machine-learning algorithm that can be applied to segment populations based on various characteristics, such as age, weight, and height. In the context of assessing nutritional status, K-means can effectively cluster individuals with similar profiles, revealing patterns and potentially identifying those who may be under-nourished. By grouping people according to their age, weight, and height, K-means facilitates the extraction of insights into distinct sub-populations, allowing for a nuanced understanding of nutritional disparities. This clustering approach may unveil clusters characterized by lower weight measures, indicating a potential risk of under nutrition. Such insights can be invaluable for public health initiatives, guiding targeted interventions and resource allocation to address specific needs within different demographic groups. K-means clustering, in this scenario, serves as a valuable tool to uncover nuanced insights into the nutritional status of diverse populations, enabling more precise and effective strategies for improving overall health outcomes.

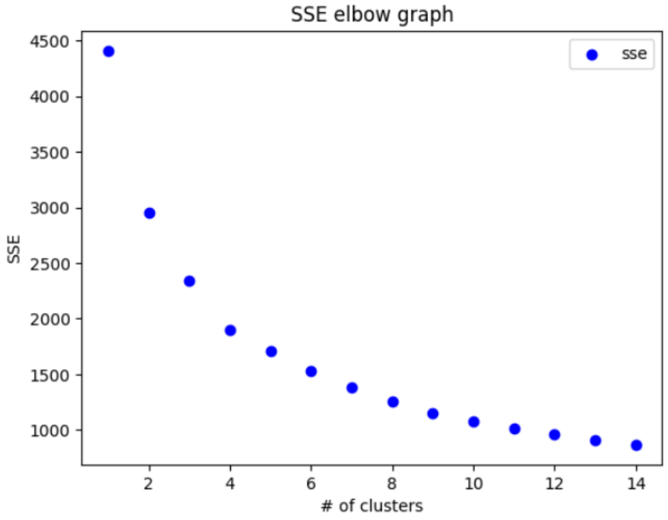


Fig. 12: SSE Elbow Graph

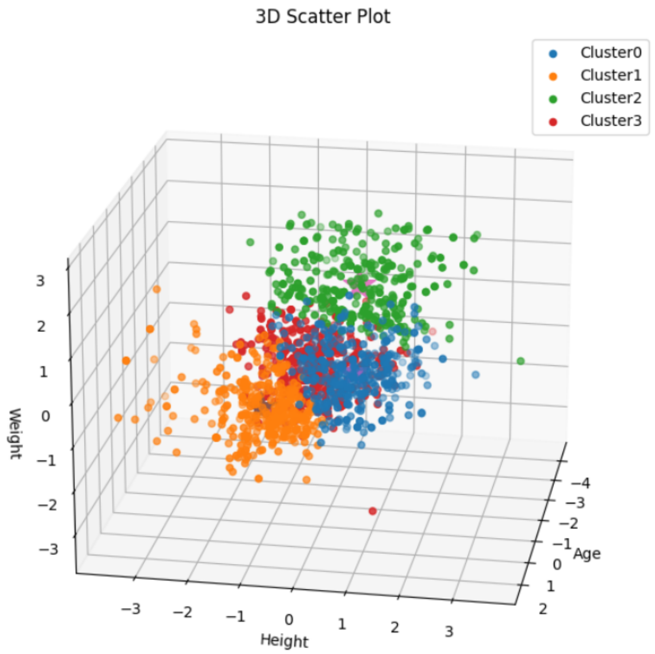


Fig. 13: 3D Clustering view 1

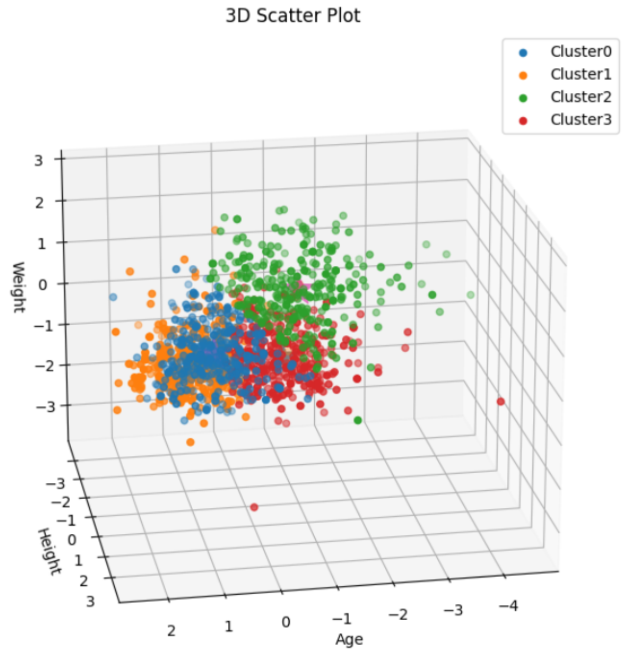


Fig. 14: 3D Clustering view 2

The SSE (Sum of Squared Errors) elbow graph illustrated in Figure 12 is a common technique used to determine the optimal number of clusters (K) for K-means clustering. The idea is to run the K-means algorithm for different values of K and calculate the sum of squared distances from each point to its assigned cluster center (inertia or within-cluster sum of squares). The SSE is then plotted against the number of clusters, and the “elbow” in the plot is identified as the point where the rate of decrease in SSE slows down. For our specific case, we pick 4 to be the ideal number of clusters and we use it to plot the K-means clustering graph shown in Figure 13 and 14.

Figures 13 and 14 illustrate the same graph from different angles. In Figure 13, it is evident that the red cluster spans the same range of heights as the green cluster. Meanwhile, Figure 14 indicates that the red cluster similarly covers the same range of ages as the green one. However, in both figures, it is apparent that clients in the red cluster weigh less than those in the green cluster. While we cannot conclusively assert statistical significance without further analysis, this observation raises a red flag, suggesting that the difference may be attributed to red clients being undernourished compared to their green counterparts

IV. Discussions

The results from SVM and Decision Tree highlight the factors with the most profound impact on the objective variable - ”DYNH Risk Score.” Given that the Decision Tree method yields higher accuracy, we can conclude that it is a superior model for our specific case and data set. SMAA can leverage this model to predict future clients' risk scores with a high level of confidence. Additionally, the model provides SMAA with insights into the most important factors influencing the risk score, guiding priority setting, and resource allocation.

The results from PCA indicate that variables in one component are closely related to one another. This implies that an increase in one variable may have a spillover effect on other variables. This information assists SMAA in prioritizing and setting goals. A secondary benefit of PCA is its ability to group similar/correlated variables together, reducing the original 95 variables to a few categories. This makes reporting and further analysis more manageable.

The clustering analysis brings together clients with similar characteristics while distinguishing dissimilar clients from each other. This aids in identifying trends or patterns unrelated to the “DYNH Risk Score”, offering additional insights. In the K-means clustering graph with k=4, it is evident that the red clusters weigh less than the green clusters, signaling a group that may be undernourished and requiring more attention and priority.

V. Limitations

The presence of empty values is not tolerable for many analysis techniques, including SVM, Decision Tree Analysis, and PCA, hence handling missing values is crucial for accurate and reliable analysis. In our dataset, consisting of 3099 rows, we've identified 2731 rows with at least one missing value, making it a significant concern. The approach we've taken involves simply removing rows with missing values. While this is a straightforward method, it has a drawback – it reduces our dataset to only 10% of its original size.

Recognizing the limitations of this approach, we are actively seeking alternative methods to address missing values while retaining as many rows as possible. One promising technique is imputation, where missing values are estimated or predicted based on the values of other variables within the dataset. Imputation enables us to preserve a larger portion of our data, which is essential for maintaining statistical power and mitigating issues such as overfitting in predictive models.

By employing advanced imputation techniques, such as mean imputation, median imputation, or more sophisticated methods like regression imputation or machine learning-based imputation, we aim to fill in missing values accurately. This not only allows us to utilize a more substantial portion of our data but also contributes to more robust and reliable analysis. It helps in reducing biases introduced by data loss and enhances the generalizability of our results.

Handling missing values effectively is pivotal for the success of our analysis, and we are committed to implementing methods that strike a balance between data completeness and accuracy in order to yield more meaningful and reliable insights.

VI. Conclusion

In summary, our data analysis of the Home Delivery Meals program at the Southern Maine Agency on Aging reveals crucial insights into demographics most in need. By leveraging advanced analytics, we've identified key factors — financial constraints, social isolation, and specific health challenges — that influence service utilization. These findings empower the SMAA to strategically allocate resources, optimizing the HDM program's impact on those who need it most. This data-driven approach not only enhances the program's efficiency but aligns with SMAA's broader mission to promote the well-being of Southern Maine's aging population. Going forward, these insights serve as a foundation for continuous improvement, ensuring the agency remains responsive to evolving community needs.