Insights From SMAA Data

Akhila Ravi and Derek Muse
Khoury College of Computer Sciences
Northeastern University

Abstract

The Southern Maine Agency on Aging (SMAA) is set up to provide resources and assistance for aging individuals in Maine. Since 2021, the organization has collected information from its clients during routine check-ins and entered the data into a large Excel sheet. The task was to explore the data and report on useful insights discovered. Statistically significant results were determined when conducting chi-square tests between clients' chronic conditions and lifestyle choices. Also, an association of the difference between Body Mass Index (BMI) of Low and High Nutritional Scores was discovered. The insight that states clients with high nutritional risk tend to decline nutrition counseling was uncovered from the provided dataset.

1. Introduction

Throughout each client check-in, SMAA employees are tasked with thoroughly surveying the health and well-being of the client. After scanning the variables included in the dataset, we decided to dive into three different areas of interest. First, we wanted to see if a client's Body Mass Index (BMI) is associated with the client's nutrition score (NSI). Next, we took a look at how the chronic conditions of a client may affect their lifestyle choices. Lastly, we analyzed the trend of nutrition scores (NSI) over time. In addition to our statistical analysis, we also decided to create a word cloud based on reviews of "Meals on Wheels", SMAA's staple food delivery service.

2. Methods and Analysis

2.1. Cleaning and formatting data

The dataset obtained was in CSV format; hence, no further formatting of the file was required. The data was read into pandas data frame since data frames provide several methods to preprocess and analyze the data. There were several missing values in the dataset. The columns that had more than 80% of data missing - ANS: Outcome: Other Referrals, ANS: Transport: If no, Transport Needed, ANS: Financial: Most Concerning Expenses were identified. Since imputing more than 80% of missing values would lead to class imbalance, these three columns were dropped/removed from the data frame i.e. excluded from further analysis. The rest of the columns with missing data were imputed using a most-frequent imputer which replaces the missing value with the most commonly occurring value from that column. The outlier detection was also performed and they were identified in the date, height, and weight columns. These outliers in height, weight, and date were excluded from the analysis.

	Minimum Value	Average Value	Maximum Value
Height (inches)	0.0	64.52	225.0
Weight (lbs)	12.0	172.38	13,540.0

Figure 1. Height and Weight information demonstrating clear outliers

The values above present the raw height and weight minimum, maximum, and averages. Clearly, data has been incorrectly entered with heights ranging from 0 inches to 225 inches (18.75 feet!), and weights ranging from 12 lbs to 13,540 lbs. This information demonstrates a clear necessity for the removal of outliers.

Before conducting T-tests between the groups, outliers are removed from the data set. The following decisions are made in filtering out the outliers: clients with weight less than 50 lbs or greater than 400 lbs are removed, and clients with height less than 36 inches or greater than 84 inches are removed. Note, these outliers were only removed for the T-tests conducted, they were not removed for the remainder of our analysis, as we believed this was simply a data entry issue.

2.2. Exploratory Demographic Data Analysis

The exploratory data analysis was performed to gain some insights and ideas about the distribution of the data. A few columns such as age, gender, eating habits, and financial status were analyzed and visualizations were performed.

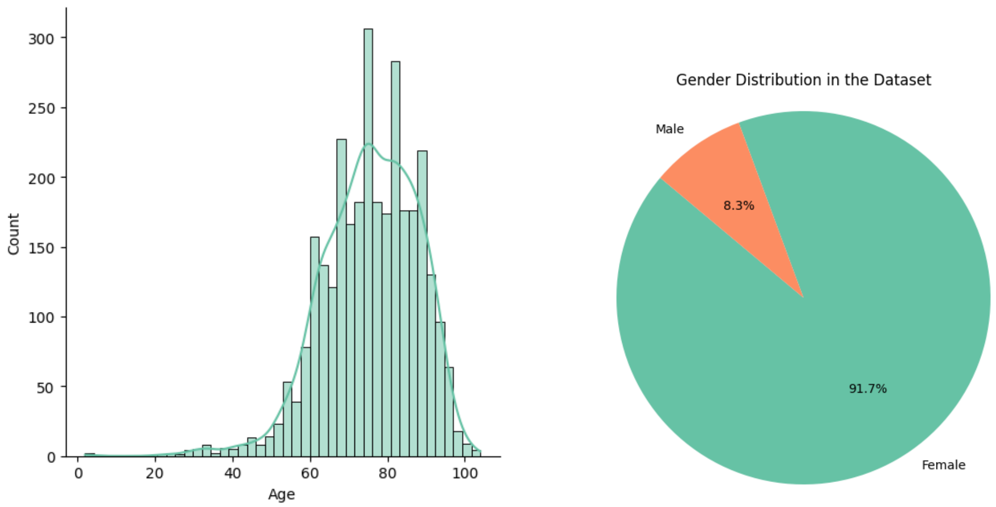


Figure 2. Age and Gender of the client base

The first plot depicts the distribution of age. From the plot, we can visualize that the majority of people in the dataset fall under the range of 65-90 years of age with the highest number of people at around 75 years of age. The second plot depicts the gender distribution in the dataset. It is surprising to note that females are the predominant gender in the dataset.

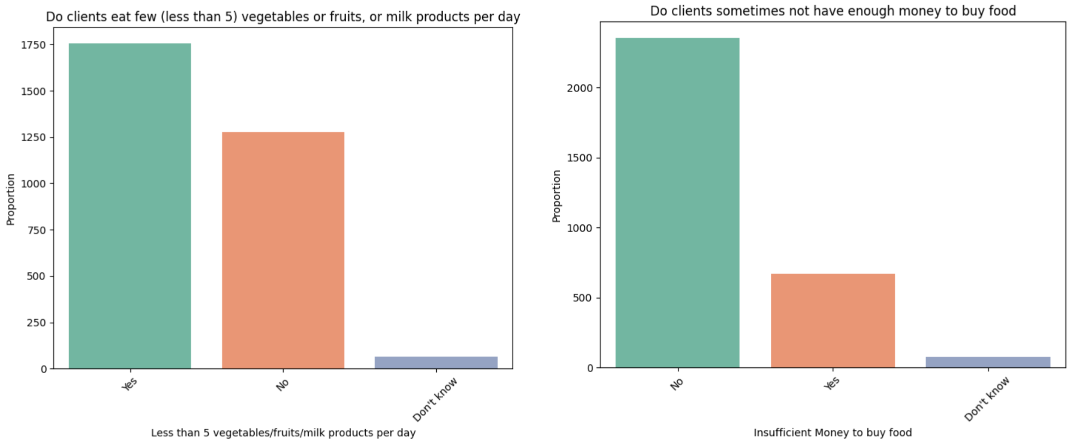


Figure 3. Vegetable, fruit and milk intake and sufficient funds to purchase food

The first plot here depicts the distribution of people who consume less than 5 vegetables/fruits/milk products. It was surprising to note that most people do not consume at least 5 vegetables/fruits daily. The second plot depicts the distribution of clients who sometimes do not have enough money to buy food. It shows that several people do have enough money to purchase food.

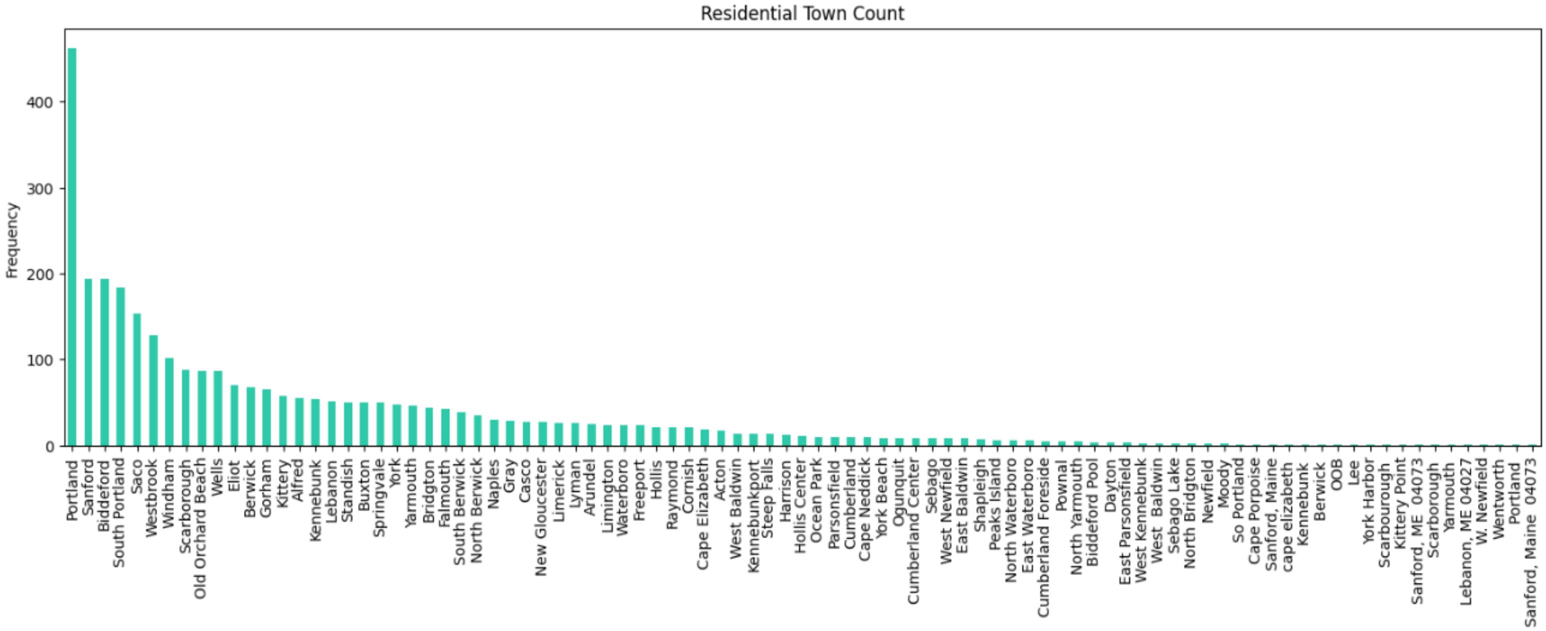


Figure 4. Client counts for each Residential Town

The bar chart above illustrates the number of client visits in each residential town since data collection began in October 2021. Note, that the count represents the number of times an SMAA employee conducts a check-in with their client, not the number of total clients in a town. If, for example, a client in Kennebunk is visited by SMAA two times in one week, the count for Kennebunk is incremented by 2. This bar chart shows that SMAA is very involved in the cities/towns of Portland, Sanford, Biddeford, South Portland, and Saco. As expected, more densely populated cities and towns have a higher count of client visits.

3. Results and Discussion

3.1. BMI Vs NSI score

The first area of exploration looks at the relationship between Body Max Index (BMI = $\frac{703*weight}{height^{2}}$) and NSI Score. There are three possible categories for NSI Score: Low, High, and Unknown. T-tests are conducted between each of the groups in hopes of finding a statistical difference between groups. Boxplots and histograms below illustrate the nature of the data.

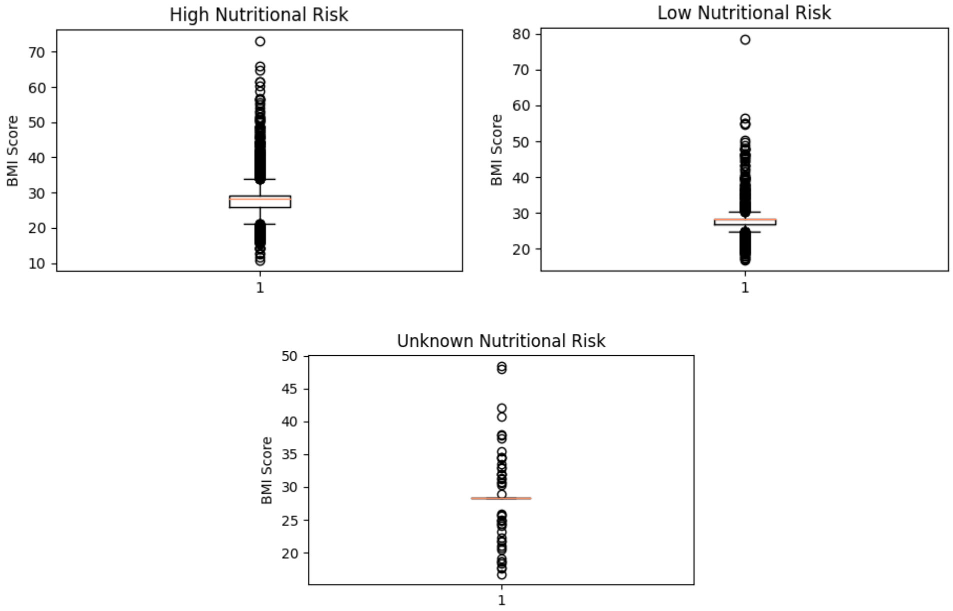


Figure 5. Boxplots of BMI Scores for each Nutritional Risk Category

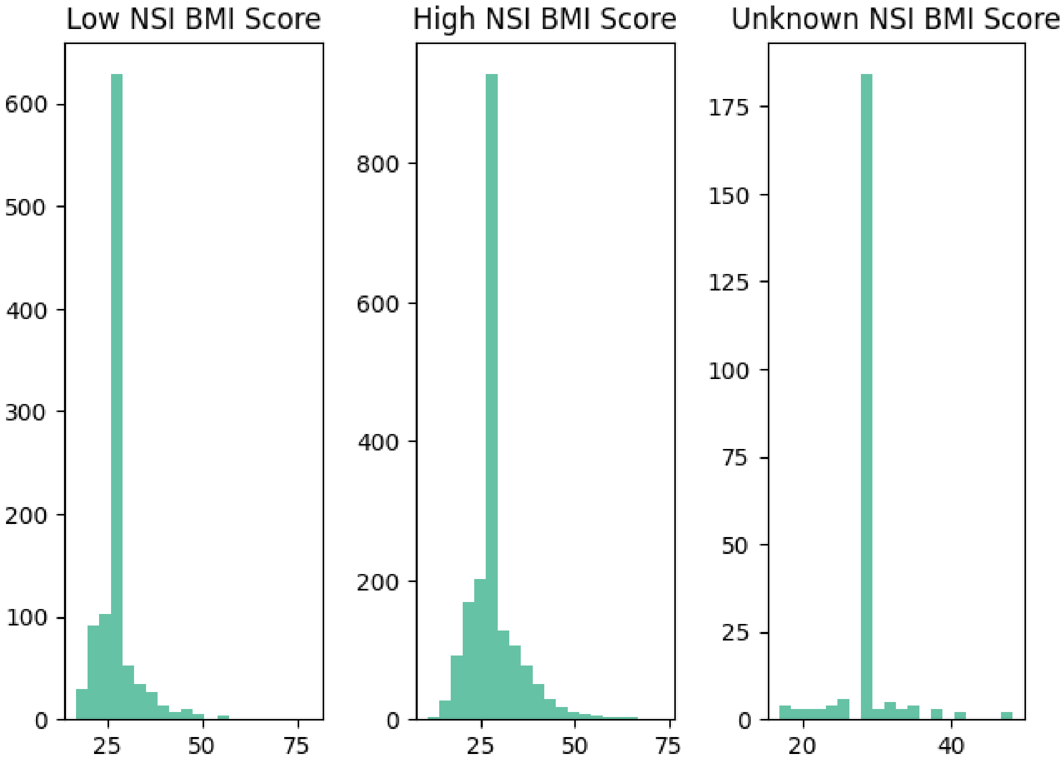


Figure 6. Counts of BMI Scores for each Nutritional Risk Category

Note, that all histograms above are centered close to the national average of 26.5. The High NSI score category appears to have higher BMI scores in comparison with the Low NSI score category. T-tests for BMI Scores between

High and Low, High and Unknown, and Low and Unknown NSI categories are conducted with the results shown below.

Groups	P-Value
High NSI vs Low NSI	0.059
High NSI vs Unknown NSI	0.427
Low NSI vs Unknown NSI	0.781

Figure 7. T-Test Results Comparing BMI Scores between different Nutritional Risk Categories

	Minimum BMI Score	Average BMI Score	Maximum BMI Score
High NSI	10.82	28.27	73.15
Low NSI	16.96	28.28	78.46
Unknown NSI	16.72	28.38	48.46

Figure 8. BMI Score Information for each Nutritional Risk Category

While we were unable to find a statistically significant result (p-value < 0.05), we found a p-value nearly less than 0.05 when comparing the BMI scores between the high and low NSI categories. When comparing averages, the Low Nutritional Risk category has a lower Mean BMI score. While we cannot conclude statistically, Low Nutritional Risk clients appear to have lower (healthier) BMI scores than High Nutritional Risk clients. The results when comparing groups to the Unknown category are not significant.

3.2. Chronic Conditions Vs Lifestyle

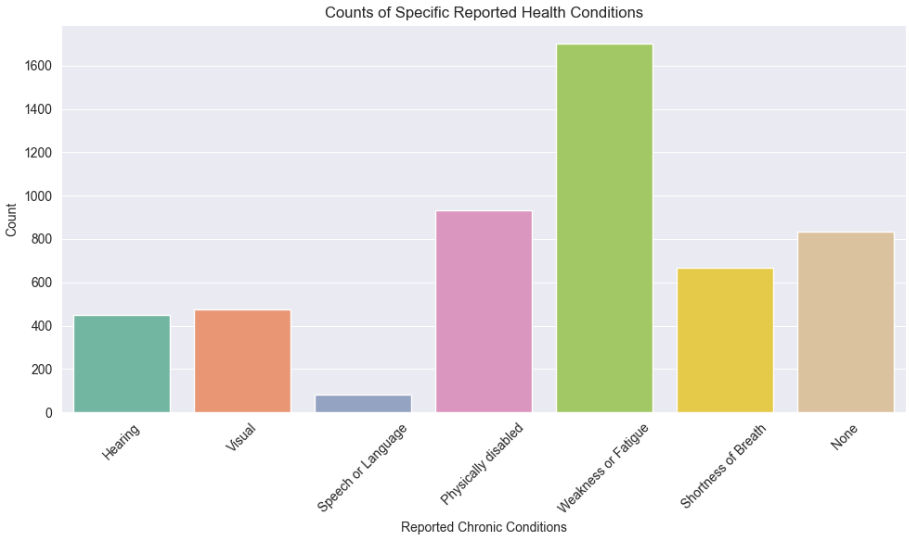


Figure 9. Counts for each Chronic Condition category

The above plot shows the distribution of reported chronic conditions in the dataset. The highest number of people suffer from weakness/fatigue as compared to other chronic conditions.

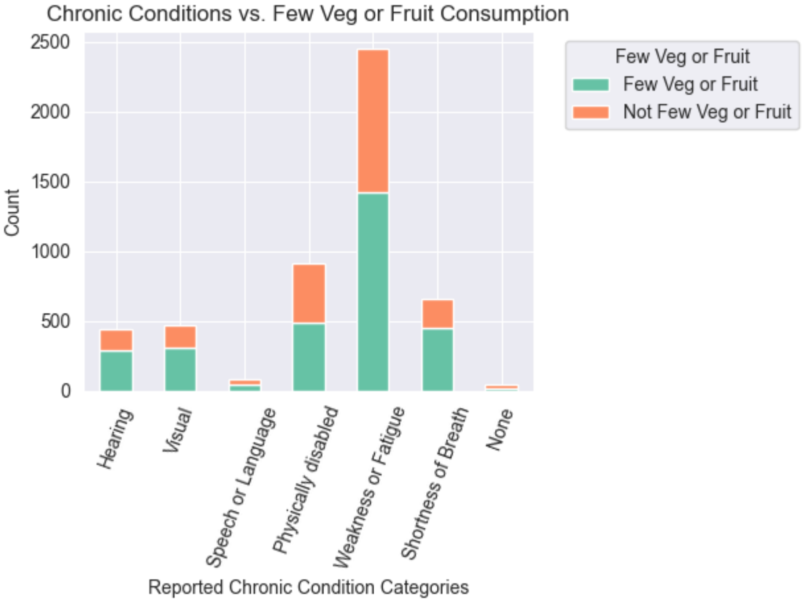


Figure 10. Vegetable and Fruit intake for each Chronic Condition category

A chi-square test was performed between the chronic conditions and consumption of fruits/veggies to determine if there is an association between them. The p-value was 2.926903744563485e-10 which suggests there is a strong association between the two columns. From the plot, it can be determined that people who consume less than 5 fruits/veggies suffer more from chronic conditions as compared to those who eat sufficient veggies.

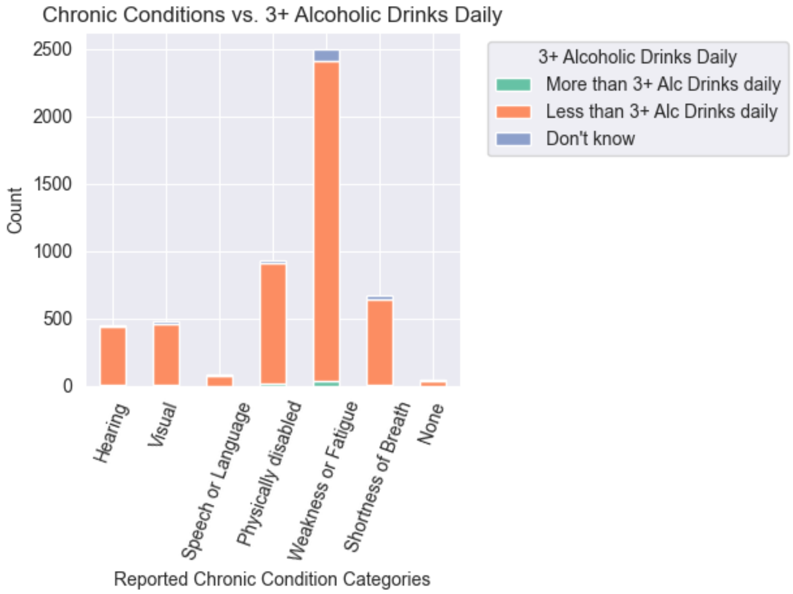


Figure 11. Alcoholic Drinks per day for each Chronic Condition category

A chi-square test between the chronic conditions and consumption of alcohol gave a p-value of 1.098243859766294e-14 which suggests there is a strong association between the two columns. From the plot, it can be determined that the majority of the people consume less than 3 alcoholic drinks and the people who suffer from weakness/fatigue are more likely to drink alcohol.

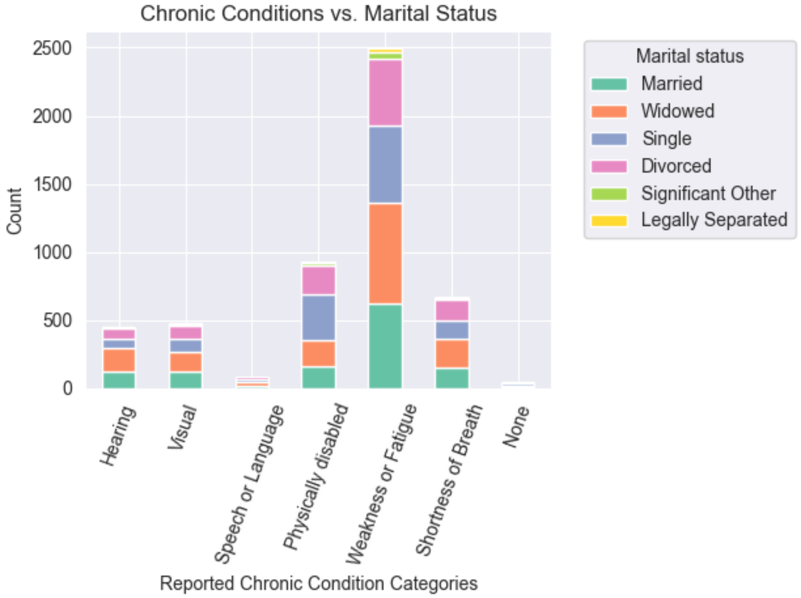


Figure 12. Marital Status for each Chronic Condition category

A chi-square test between the chronic conditions and marital status gave a p-value of 7.113499606169692e-10 which suggests there is a strong association between the two columns. The plot does not show any significant insights despite the p-value of the chi-square test being close to zero.

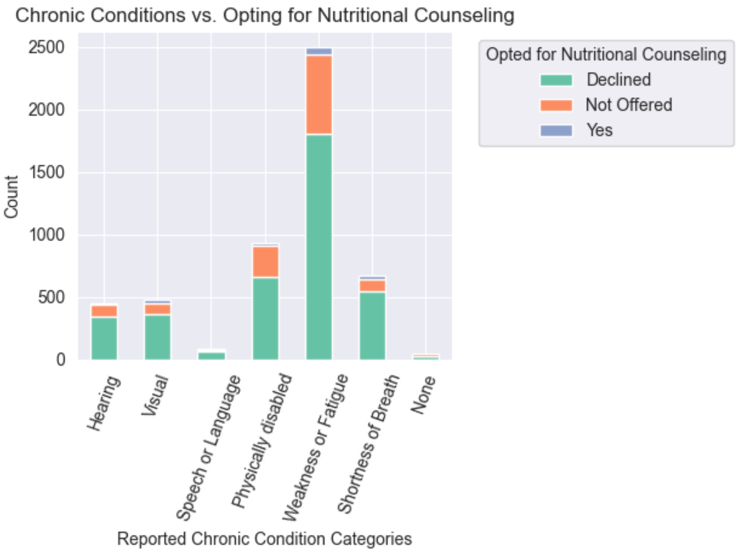


Figure 13. Nutritional Counseling responses for each Chronic Condition category

A chi-square test between the chronic conditions and opting for nutritional counseling gave a p-value of 6.122491822335625e-30 which suggests there is a very strong association between the two columns. The plot does show that people with chronic conditions are more likely to decline nutritional counseling than the ones who do not suffer from chronic conditions.

To determine if any columns do not bear any association with chronic conditions, a chi-square test was performed between chronic conditions and several other columns. The columns that did not have any association with chronic conditions were gender, veteran status, homebound, and eating alone as their p-values were high which indicates strong dissociation.

3.3. Is there any change in the NSI score?

The next question to be answered was to determine if there is any change in the NSI score over the two years and find insights from the NSI score vs nutritional counseling.

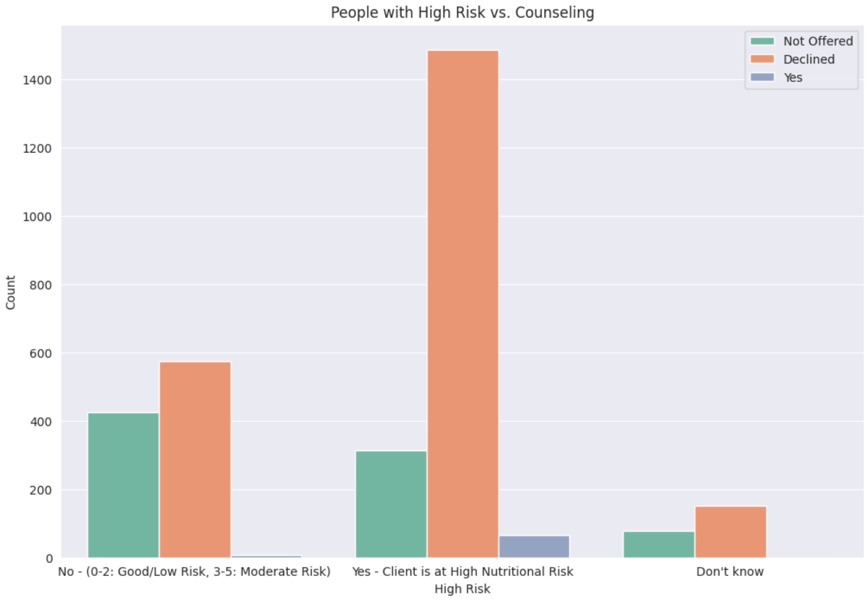


Figure 14. Counseling responses for each Nutritional Risk Category

This plot shows the distribution of NSI scores vs opting for nutritional counseling. It was surprising to note that people with high nutritional risk decline nutritional counseling the most.

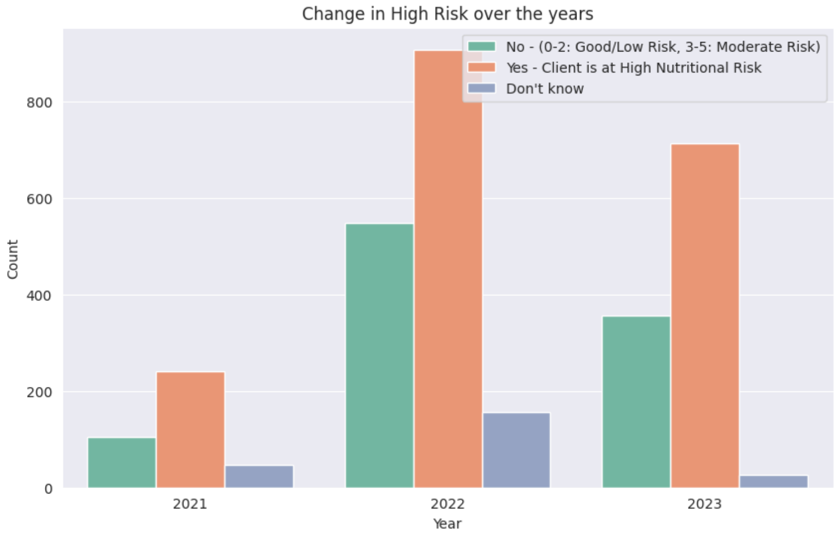


Figure 15. Nutritional Risk Counts from 2021, 2022 and 2023

Also, there weren't any significant changes in the count of high nutritional risk over the years since the data available is from October 2021 which explains the imbalance in the dataset.

3.4. What do the users say about Meals On Wheels?

The reviews about Meals On Wheels were extracted from Influenster and Indeed. Once parsed into a text file, the text was preprocessed to remove the non-alphabetic characters and removal of stopwords. Once the preprocessing was done, a word cloud was used to visualize the high-frequency words and the overall sentiment of the reviews.

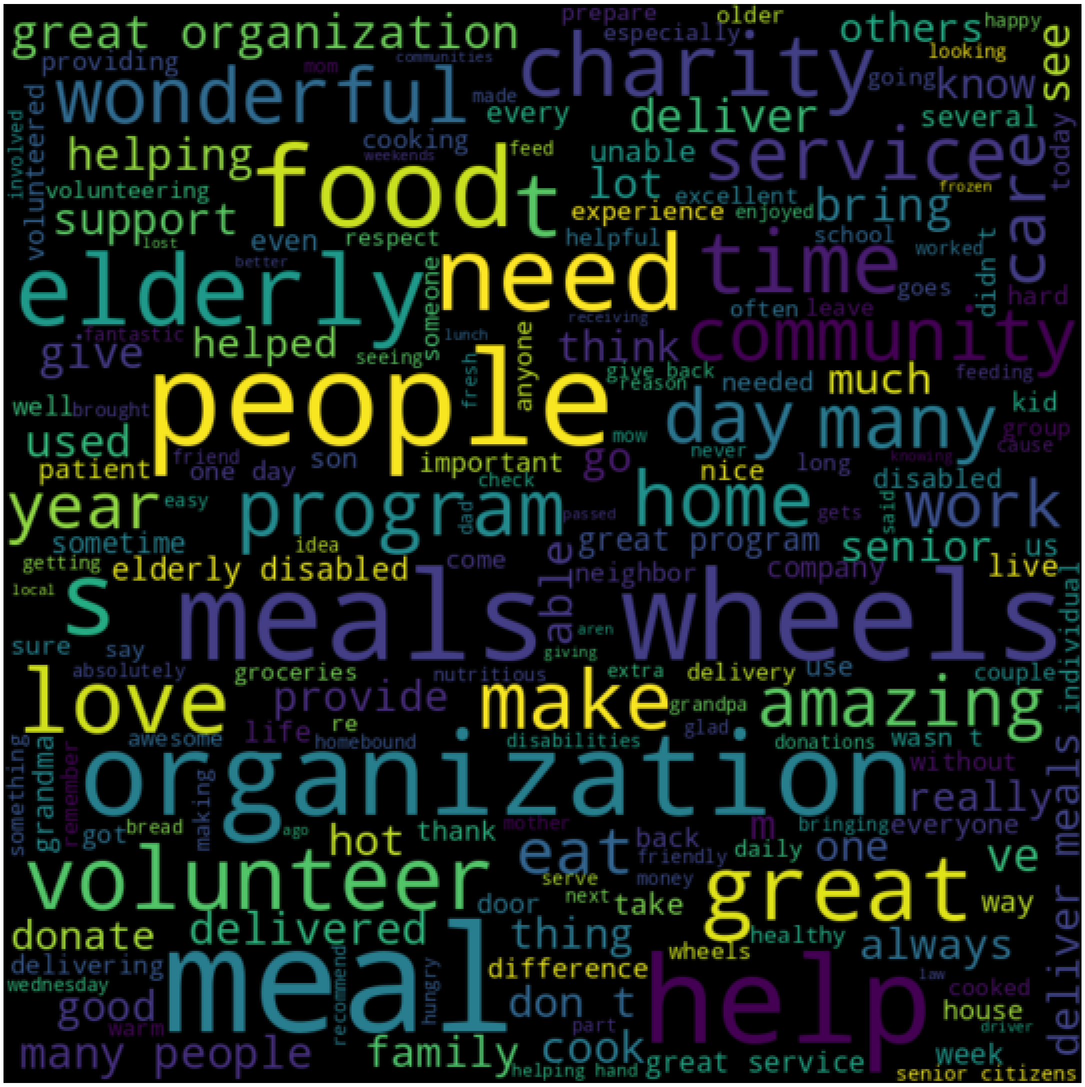


Figure 16. Word Cloud of frequently used words from Influenster and Indeed Reviews

The word cloud depicts a positive sentiment and highlights the important and frequently occurring words that describe the user's sentiment about Meals on Wheels. It highlights the words meals, great, volunteer, wonderful, and charity which indicates that users feel positively about this program.

4. Discussion

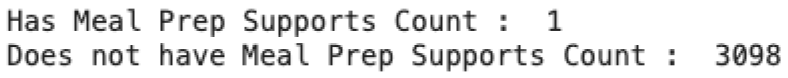
A T-test performed between NSI Score and BMI did not yield a significant result, but it appears as though Low Nutritional Risk Clients have lower BMI. This takeaway confirms the success of SMAA's ability to evaluate the nutritional status of their clients.

A chi-square test performed between chronic conditions and several other columns to identify association yielded interesting results. The columns that bore strong association were consumption of fruits/veggies, consumption of alcohol, marital status, and opting for nutritional counseling. There was a strong dissociation between chronic conditions and gender, veteran status, eating alone, and being homebound.

The NSI score trend revealed that people who are at high nutritional risk are the ones who decline nutritional counseling more. There weren't any significant changes in the count of high nutritional risk over the two years of data we have available. The word cloud depicted positive sentiment towards the Meals on Wheels program and displayed the high-frequency words used to describe their reviews about the program.

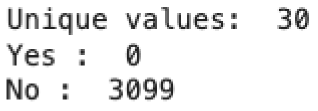
5. Limitations

While missing values and outliers were obstacles we had to overcome in conducting the analysis, we were limited to only one of our desired areas of exploration. We hoped to explore the relationship between a client's ability to prepare meals ('Eligibility2: Meal Prep Ability') and that client's meal support status ('Eligibility2: Has Meal Prep Supports'). We discovered the following:



```
Has Meal Prep Supports Count :  1
Does not have Meal Prep Supports Count :  3098
```

Every client except for 1 does not have meal support unless the column has been misunderstood (in which case every client does have meal prep support). On top of this, the Meal Prep Ability column contains 30 unique values, none of which describe an individual able to prepare a nutritious meal successfully. The lack of variability in these two columns led to the dropping of this area of exploration from this report.



```
Unique values:  30
Yes :  0
No :  3099
```

6. References

1. https://www.influenster.com/reviews/meals-on-wheels/reviews
2. https://www.indeed.com/cmp/Meals-On-Wheels/reviews
3. https://matplotlib.org/stable/index.html
4. https://www.smaaa.org/
5. https://www.cdc.gov/nchs/data/nhanes/databriefs/adultweight.pdf