# Unsupervised Wisdom Executive Summary, Mike Goldweber

## Key Findings

There were several things that stood out. First, women were 66% of the injured. The majority of injuries occurred at home, and while the greatest number of diagnosed injuries were ‘fractures’ at 7K+. The single problem for both females and males were ‘internal injuries’ to the ‘head’ (17K total) (Table 1). Connect to the injuries is the single biggest product connection was to ‘floor or flooring material’. Also, I discovered that there was a significant increase of ER visits in 2022, compared to the previous 3 years (Fig 1). Age, race, Hispanic surprisingly didn’t have a significant impact on the distribution of injuries (Fig 3).

## Summary of Project Approach

I decided early on there would be benefits to thorough data exploration. As a result, I decided to solely use on the primary dataset provide by the contest. My thought was that the categorical information would provide context for the falling injuries, which the narrative would provide insights into the reasons behind the falls. As a result, I looked at each column individually, followed by looking at combinations of the columns to find the combination(s) of categories from the key columns that would expose the problem. Admittedly, the approach was tedious and requires some patience; but by identifying the main factors behind the injuries, I believe an approach could be developed for mitigating these injuries across the elderly population.

One interesting approach to this project was the use of SQL on the dataset to explore these combinations of data categories. I believe the data exploration was enhanced, and it allowed me to find the biggest category combinations behind the injuries. For example, the mosaic plot shown in Figure 2 demonstrates the combination of the sex, diagnosis, location, and body part features of the dataset.

One problem was the significant number of unknown and unstated items in many of the data’s features. This made it difficult to be certain that race and ethnicity was not a part of this problem, and it presents uncertainty with the location. I firmly believe many of the unknowns in the location column are home injuries, so I am sticking with this as a focus to any solution.

I used a *Partitioning Around Medoid* model because it works with a mix of continuous and categorical data[[1]](#footnote-1). The model’s results I believe confirm what was discovered in the exploration. First, I split the set into two groups, one that covers 2019-2021, the second covers 2022. Initially this was due to model size limitations (max set of 65K). However, this was an unexpected opportunity to see if some insights into the higher injury counts could be discovered. The 2019-2021 PAM model shows a lot of similarities between the clusters (Table 2), and a distribution of the injuries and location. However, the 2022 PAM model shows one major cluster dominating this portion of the dataset. I was surprised to see that neither model focused on the injuries or location. In both sets, cluster 1 was heavy with the fracture and internal injuries.

A major limitation of my analysis is the failure to get aid from an LLM to analyze the narrative data. This was due to technical problems, rather than ignoring this aspect of the project.

## A graph of injury by day Description automatically generatedData Exploration Visualization

Figure - Showing the frequency of injuries by day. Note the last quarter (2022) shows a major increase of daily injuries.

A white text with black text

Description automatically generated with medium confidence

A pie chart with numbers and text

Description automatically generatedA blue and red rectangles

Description automatically generated

Table - Showing the top 10 frequency of top injuries broken down by gender.

Figure - This plot show all of the key elements of the data exploration brought together, sex, diagnosis, locations, and body parts.

Figure - Breakdown of Pop by Race. N.S group spoils the possibity of being sure if race is a factor in the injuries.

## Modeling Visualizations

A white background with numbers

Description automatically generatedA screenshot of a computer

Description automatically generated2019-2021 On the left, 2022 model data on the right.

Table - 2022 Clusters

A blue rectangular shapes with black lines

Description automatically generated with medium confidenceA diagram of a graph

Description automatically generated with medium confidenceA graph of a plot

Description automatically generated with medium confidenceBlue squares with black lines

Description automatically generated

Table - 2019-2021 Clusters

Figure 8 - Cluster and Injury factors from the PAM model for 2022.

Figure - 2019-2021 Clusters with sihouette of 0.28.

Figure -2022 Clusters. Silhoutte of 0.26

Figure - Cluster and Injury factors from the PAM model for 2019 - 2021.

1. Preud’homme, G., Duarte, K., Dalleau, K. *et al.* Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. *Sci Rep* **11**, 4202 (2021). https://doi.org/10.1038/s41598-021-83340-8 [↑](#footnote-ref-1)