We conducted data engineering before uploading the OHAS (Occupational Health and Safety) CSV dataset into Tableau and found no duplicates or nulls but some missing values, as shown below.

A screenshot of a computer code

Description automatically generated

After the cleaning, our dataset had 2,114 instead of 2,129 rows and the same 13 columns.

A screenshot of a graph

Description automatically generated

Our decision to retain 17 out of 48 common diseases for further data analysis resulted from a collaborative discussion. The filtered dataset we arrived at had no duplicates or nulls and only one missing value, as shown below.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

Our filtered dataset consists of 218 instead of 2,114 rows and the same 13 columns.

We are analyzing 17 diseases and 130 symptoms categorical unique values. The average height ranged from 62 inches to 76, the average weight from 92 to 600 pounds, and the average BMI from 17.1 to 53. The age ranged from 16 to 57 years old. We had an imbalanced dataset when referring to 97 males and 121 females. For clarification, the columns Disease and Symptoms CUIs relate to the disease Concept of Unique Identifier in the Unified Medical Language System (UMLS).

We created three dashboards in Tableau. The first dashboard, Symptoms and Diseases Overview contains two bar charts. The bar charts display the types and count of diseases, and the most prevalent symptoms by disease type. The second dashboard, Regional Disease Trends, consists of a stacked bar chart and a circle chart. The stacked bar chart includes the count of diseases by fourth regions: northwest, northeast, Southwest, and Southeast. The circle charts display the number of diseases by region and gender. The third dashboard, Severity Factors Comparison, displays a scatter plot, and a box plot. The box and scatter plot displays high, low, and medium disease severity classifications. The scatter plot compares the severity of the diseases by BMI and weight. The box plot establishes the gender differences by age in each disease severity.

**Results:**

Hypothyroidism emerged as the most predominant disease, with the most at 23, while migraine was the least, at 7. The symptoms of shortness of breath, asthenia, and fever were the most prevalent.

The Southeast region exhibited the highest disease prevalence, with 75 diseases, while the Southwest region had a relatively lower prevalence, with 44 diseases. Notably, chronic diseases were the most predominant, affecting 95 out of 218 individuals, underscoring the need for region-specific healthcare strategies. Males suffered more chronic diseases in all regions, males at 52 and females at 43, except in the Southwest, females at 11 and males at 9.

The impact of health metrics and demographic factors on disease severity was also notable. BMI and weight were found to affect disease severity, with higher values correlating to higher severity. The higher the BMI, the individual is either overweight (BMI 25-29.9) or obese (BMI 30 or higher), the higher the disease severity. Males were found to suffer from a higher severity of diseases than females, more notably in younger males. These findings highlight the importance of health metrics and demographic factors in understanding disease severity.

**Conclusions:**

The prevalence of diseases varied across regions, with chronic diseases being predominant, emphasizing the importance of region-specific healthcare strategies.

Gender differences in disease prevalence and severity were observed, with males generally experiencing higher severity, particularly noticeable in younger age groups.

Works Cited

Disease Prediction Using Machine Learning

<https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning?resource=download>

Disease Prediction through Symptoms (Tableau Data set)

<https://www.kaggle.com/datasets/usamag123/disease-prediction-through-symptoms/code>

IF [Latitude] >= 35 AND [Longitude] <= -90 THEN "Southeast"

ELSEIF [Latitude] >= 35 AND [Longitude] > -90 THEN "Northeast"

ELSEIF [Latitude] < 35 AND [Longitude] <= -90 THEN "Southwest"

ELSE "Northwest"

END

The dataset contains diseases with specific symptoms. Symptoms are indicated by 0 and 1. Value 1 indicates symptoms are available, and 0 indicates no symptoms. This dataset can predict the accuracy of several machine learning models like SVM, Naive Bayes, Decision Tree, Random Forest, etc.

Target disease? But cannot change to 1 or 0 prognosis