ETL PROJECT BREAK DOWN

Michelle Gillis & Raquel Alto

We sourced some Alzheimer’s clinical data from an unspecified organization on Kaggle and a nutrition study dataset from the CDC. Both were thoroughly cleaned using a variety of methods, starting with the removal of unneeded columns. Three of the original seven columns in the Alzheimer’s dataset were removed. This left the ‘mmse’ (Mini-Mental State Examination), ‘ageAtEntry’, ‘cdr’ (Clinical Dementia Rating), and ‘memory’ columns. No rows needed to be eliminated, so next the ‘ageAtEntry’ column was binned into age groups while the remaining three columns were aggregated to find the mean of their respective values. The data sourced from the CDC was more difficult to work with. There were many empty rows and mixed categories needed to be eliminated in order to reduce the data to age group weight. Many of the age groups listed were not part of the target demographic investigated, so those were removed as well. The result was one table showing five age groups alongside average cdr and age of the individuals for the group, and another table showing age group and average frequency of obesity. These cleaned tables were exported into PgAdmin through a python engine connection and then joined by age group. The final result is a table demonstrating the average occurrence of obesity, mean age, and mean cdr for dementia-prone/dementia-confirmed age groups.