Team Puebla: Lily Gates, Nainika Kadabha, Theodore Rose, Andre Soriano, Anna Wendt

Final Project Report

I. Describe Dataset and Database

Our team's dataset and database centers around the "Alzheimer's Disease and Healthy Aging" dataset. The data for our project was provided by the Centers for Disease Control (CDC) and Prevention Division of Population Health. We selected this dataset due to our common interest in exploring the factors that contribute to and influence cognitive decline, which is a main element included in our project. Our dataset aims to find correlations between these factors to infer and analyze healthy strategies we can adopt to improve our aging prognosis. Our database includes certain demographic and lifestyle factors of individuals that can be connected to Alzheimer's and cognitive decline, such as age, race, sleep, physical activity, overall healthcare, location, and more. We included questions and queries that related about what caused and correlated with Alzheimer's development. Our dataset consists of 38 columns and over 250,000 rows. Some of the columns consist of the year, location, category of question, question itself, data value, and demographic groupings, The questions and queries are grouped into seven different categories, which include cognitive decline, mental health, physical health, overall health, smoking and alcohol usage, screening and vaccines, and caregiving information.

The objective and primary goal of our dataset and database is to deepen the understanding of Alzheimer's, including the risk factors associated with the disease, enhance patient care, inform public health strategies, and ultimately contribute to finding Alzheimer's prevention tactics and potential cures. Our dataset will provide useful information to a

variety of entities, including health-conscious individuals and healthcare professionals. Health-conscious individuals can use this data to make decisions that benefit their cognitive health and longevity. Health researchers and medical scientists can use this data to uncover potential causes such as genetic or environmental factors, and propose new treatments. Public health analysts can use the dataset to track the frequency of Alzheimer's and cognitive decline across different demographics, regions, and age ranges, identifying high-risk populations and informing public health interventions to bring awareness to those affected. Government and healthcare systems can use this data to allocate resources optimally and possibly design programs following trends in the data to support those at risk of Alzheimer's.

II. Changes Since the Initial Proposal

Since our initial proposal, we have adjusted our project in a few ways. Firstly, we originally planned to reduce our 250,000 rows significantly. However, after speaking with our teaching assistant, we were informed that going over the row count suggestion is acceptable, especially if this data is useful to our database and objective.

We have adjusted the scope and focus of our project in a few aspects, but there were no significant changes. We originally were a bit unsure of what to focus on. We were interested in studying factors that correlated with Alzheimer's and cognitive decline or potential treatments to reduce Alzheimer's. As we moved further into the project, it became apparent to us that focusing on demographic and lifestyle factors would be easier to find concrete data on, and to build queries off of. Choosing this, and highlighting these lifestyle factors and their correlation to cognitive decline, allows us to use our database to inform individuals on which healthy lifestyle choices may be associated with protecting cognitive function.

We reduced four different entities and attributes since our initial proposal. There were some individual entities and attributes we did not include in our final database. The common reasons for removing these attributes was that they contained data that we deemed would not be useful to our project, the data indexing could be clearer, or to eliminate redundancy. These entities include RowId, Datasource, Response, Data_Value_Alt, ClassId, TopicId, QuestionId, LocationId, ResponseId, Data_Value_Footnote_Symbol, Sample_Size, Stratification Categories, StratificationId3, and certain years. There were also a few categories of data that we chose to simplify or remove. Firstly, we reduced the locations used. There were fifty-nine locations, including fifty states, four territories, four regions (North, South, East, and West), and an overall location. We simplified our database to only include the regions and overall location, largely for organizational and simplification purposes. Additionally, we reduced the years used. We had initially decided to reduce our dataset from 2019 to 2021, but we noticed something unusual about the data we used. There was a survey that asked the same questions for three years in a row (2019, 2020, and 2021), but only to certain locations. We concluded that using 2021 would provide more accurate, concrete data, because it was the only year in which they stored survey responses for the regional and total locations. Moreover, we reduced the number of questions and queries asked in our database. We wanted to include a wide range of surveying factors in our database that have an associated impact on Alzheimer's development, and aimed to have our questions correlate with all of the locations that we included for comparison purposes. While it was disappointing, we chose to exclude questions 6, 7, 8, and 9, due to the fact that the locations

we used did not survey populations on these questions. Lastly, we chose to reduce the demographics and age groupings used in our database. We originally planned to use ages 50_to_64, 65_and_older, and age_overall as our age groupings. In our final database, we ended up only using age_overall to reduce the number of datapoints. As for demographics, we wanted to highlight inclusivity, and make sure that we were including data for a wide range of demographics, and prevent cutting out any specific races and genders. While we did consider how we could reduce these data points for simplicity purposes, diversity was a key point to allow this dataset to be useful to a wide variety of individuals and entities, so we made the decision to use the original demographic groups. It is interesting to note that this data was collected based on individual demographic categories and characteristics, not based on their intersections. For example, a question on measuring the prevalence of obesity only measured female and white, not if an individual was female and white.

We have envisioned some new questions that our database can answer, or build off of. It is important to note that our database could be thoroughly expanded, as there are several lifestyle habits and data points that may be associated with Alzheimer's and cognitive decline. Some of these are more complicated to include, or lack sufficient data to truly analyze their association. We did include questions about smoking and alcohol usage, but this substance usage category could be expanded to include the presence of other drugs. Additionally, this category could expand into whether or not the substance usage was previous, infrequent, recreational, or habitual, rather than just a general analysis on current usage. Moreover, we do include questions regarding staying up-to-date with medical visits and vaccines, but we do not include medications. Including questions about medications could take two paths- analyzing what medications individuals take and if or how they correlate to cognitive decline, or if individuals are taking the medications that are prescribed

to them and following medical advice. Additionally, we included questions relating to depression diagnosis and frequent mental distress, but these could be expanded to include other mental health categories and conditions: such as anxiety, post-traumatic stress disorder, bipolar disorder, and so on. Different conditions have different correlations and associations to Alzheimer's and cognitive decline, and if we were to expand our database and its scope, this would be an interesting point to touch on. On a similar note, the health conditions we included, such as diabetes and high blood pressure, could certainly be expanded to include a wide variety of health conditions and their proposed association to Alzheimer's and cognitive decline.

III. Project Evolution

Since our Progress Report, we have updated our dataset to include more stored programs. We added views, functions, and procedures to our dataset. We included three views in our dataset, with the overall intention to create simplified, table-like codes that are easy to refer back to. We were inspired to create these to allow users to have easier access to compiled information in our database, and gain some further insight on some of our queries. Additionally, they act to further analyze some of our individual questions regarding our information. Our first view is of all of the distinct questions asked, regions studied, and the corresponding demographics asked. The purpose of this first view was to amplify what questions were asked to specific demographic groups. This view displays the year the questions were asked, the location surveyed, the question category and topic, the question itself, and the demographics surveyed. This view provides great data for anyone new to using

this database, allowing them to explore the questions asked, which can help tailor their further queries. Our second view focuses on the maximum, minimum and average data values in the overall U.S. region for each question. The purpose of this view was to answer what the minimum, maximum, and average data values for each question in the United States were as a whole in 2021. This view calculates and displays the minimum, maximum, and average data values for each question. Our third view looks into demographics with a high prevalence of negative health indicators. The purpose of this view was to answer which categories exhibit negative health indicators for the majority of demographic groups. This view displays the demographic groups in which the average data value is greater than 50% for negative health indicator questions. As for our procedures and functions, we had 2 of each. Our first procedure compares race by topic. Our procedure takes the input of a topic, such as "Prevalence of Sufficient Sleep", and produces a table of the average_value with low and high confidences and numbers of data points by race. This allows the user to examine the differences in specific conditions between specific races to determine possible outliers and any other observations. Our second procedure is titled topic by region. This procedure allows for a user to input a region and get the topics and averages for that region in descending order. Moving onto our functions, our first function looks into the percentage of smokers above 50. This function allows the user to input a region and get the percentage of respondents that answered 50 or over when asked if they have smoked more than 100 times in the past and are currently smoking. Our second function looks into the second highest reported topic for regions and race. This function allows the user to input a region and a race to obtain the second highest reported topic for both of the specified categories. We decided to do the second highest as the highest among the large majority of both regions and races was High Blood Pressure. We added all of these new views, procedures, and functions to allow

our users and ourselves to have more insight into our database, and to answer some of the unanswered questions and intersections we had about our data.

IV. Future Improvements

Of course, we were limited in time and resources to complete our database and project. Our dataset goes into a significant amount of detail, but Alzheimer's and Healthy Aging is such a large topic to be analyzed, with much more information being available to elaborate on. If we had unlimited time, we would have included more elements, categories, and queries in our dataset. This would have allowed us to dig deeper into our topic, and provide more specific information. Our dataset included 40 detailed questions, but this section could easily be elaborated on. Some of these elaborations could have related to deeper past medical history, more mental health conditions, specific physical activities and routines, education, income, diet, and familial-history. Additionally, we originally limited our data to be from 2019-2021, and then chose to just use 2021. With more time, we could have expanded this time frame. We only included regions and overall locations, being states and territories. With more time and hands, this could have been expanded, either to smaller, more specific locations, or to a worldwide scale. Lastly, with more technical expertise, we ideally would have liked to create more views. As a team, we struggled with forming and finishing our views. In a perfect world, we would have been able to create these and more with no problems.

V. Team Contributions

We split the tasks for our Progress Report evenly, with our project leader, Lily, assigning tasks to our team members based on their individual strengths reported at the beginning of our project development. From the beginning, Lily listed our team members strengths, allowing her to assign us tasks that best reflect our strong points, which has been extremely helpful throughout this project to ensure that everyone completes their best work and is confident in their final work produced. For the Project Report, Theo took on the ERD portion of the project, making sure to update it according to the TA's comments and suggestions, and also added two functions and two procedures Lily completed the Excel portion and views, Nainika completed the Database Export file section and assisted with the views and functions, and I (Anna) completed the Written Progress Report section. The team worked cohesively and collaboratively to stay on top of progress, due dates, and to answer any questions other team members had about their sections or other sections of the project.