Final Project Report

Group 2
Mia Manning, Erick Salmeron, Victoria Stavish, Daniel Trinh, Kamari Trotz
INST327 (0101): Database Design and Modeling
Dr. Vedat Diker and Professor Pamela Duffy
December 12, 2022

Introduction

City trees and green space–in addition to providing many other ecosystem services–increase biodiversity, offer shelter from the hostile urban environment for animals whose habitats have been fragmented by development, mitigate the urban heat island effect, keep the air clean, and sequester carbon (McCoy et al., 2022). Additionally, the existence of green space presents the opportunity for local and state governments to implement unique public health initiatives to combat disparate health outcomes. Biodiversity affects not only the health of the natural environment but of the people interacting with that natural environment. Public health initiatives that prescribe green space exploration would offer an alternative to prescription medication. A more holistic approach to managing chronic health issues such as depression, high blood pressure, and diabetes could benefit from an understanding of who has access to green space, and where that green space is located. This is where our database would be most useful.

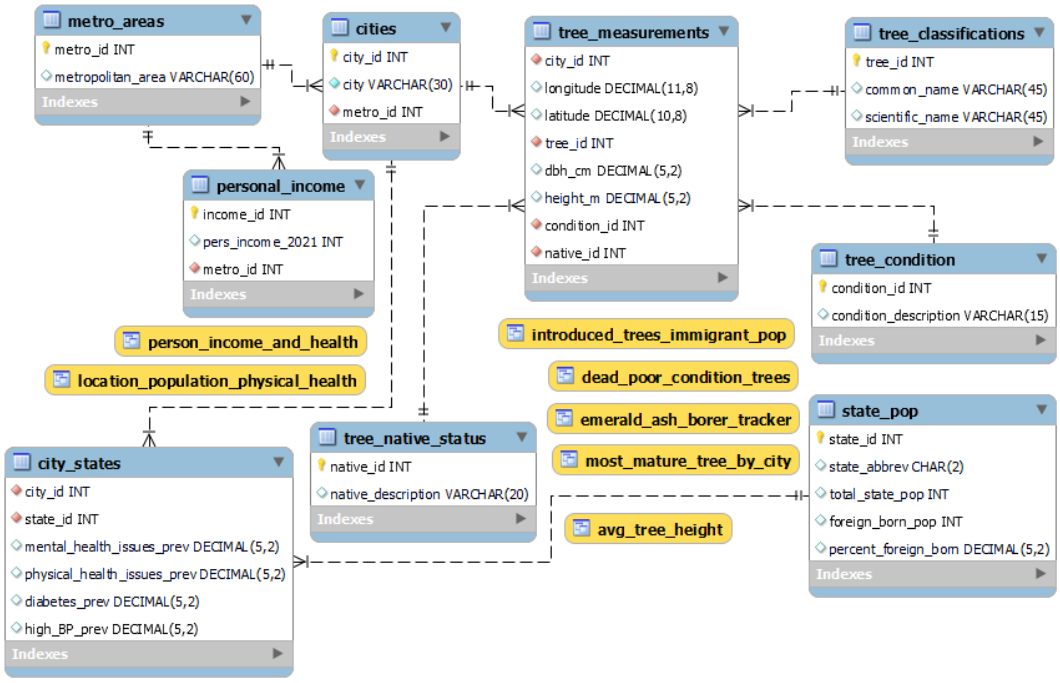
We wanted to create a database that inventories existing tree specimens, but also balances that with consideration of the surrounding demographics and what they might desire from the natural environment. For example, they might want species planted that remind them of their home country, or plant species that might be easier to maintain but are non-native or introduced. We wanted to relate a tree inventory to the following concerns: biodiversity, the potential influence of green space on health outcomes, the abundance of trees based on the wealth of a population, and the occurrence of non-native tree species stemming from a large immigrant population in a state or city. Ultimately, the relationship between existing trees and the surrounding demographic is what our database is modeled after.

Database Description

The purpose of our database is to provide information about trees in major U.S. cities while also providing context about the city such as income, population, and health data. The heart of our database is a tree inventory which pulls 25 entries each from the following 15 city datasets: Baltimore, MD; Columbus, OH; Seattle, WA; Tampa, FL; Durham, NC; San Jose, CA; Honolulu, HI; Aurora, CO; Austin, TX; Washington, DC; Sioux Falls, SD; New York, NY; Albuquerque, NM; Knoxville, TN and Louisville, KY. Our database consists of eight tables. We augmented the base tree inventory tables using entities with select attributes from flat data collected by the U.S. Census, Dryad, the Bureau of Economic Analysis[BEA], and the Centers for Disease Control[CDC].

Logical Design

Figure 1. *City Data and Tree Inventory Final Database ERD*



We organized our database around the cities table and by extension the tree_measurements table; each major linking table stems from the 15 major US cities we selected. We felt it was more intuitive that users utilize the cities and city_states tables as the gateway to demographic data, and use the tree_measurements table as the bridge to tree specific data. The tree_classifications, tree_native_status, and tree_condition tables are self-contained as they constitute a small range of values that would be duplicated in the base measurements table if not separated out. Their primary keys are foreign keys in the tree_measurements table in order to establish a strong relationship between each entity.

Next, we determined it best that all primary keys were auto-incremented. That way any newly inventoried trees or demographic data inserted into the database is numbered in accordance with the existing system. We also felt the city column in the cities table should be non-null, as any queries of our database are heavily reliant upon the city name. Lastly, we specified that scientific_name in the tree_classifications table should be unique–especially since the purpose of the genus and species naming convention is to designate unique species diversity.

Physical Database

Our goal for this database is to provide data that offers potential insight into *why* certain trees are more abundant in certain cities, *how* income in each city might correlate with trees in better condition, and whether cities with a larger concentration of trees in good condition correlates to better health outcomes. Data retrieved from Dryad influenced what additional data sets we consulted. Tree inventory data became the most important aspect of our database, which relied on cities. Data from other datasets had to be linked to our tree inventory based on the city, the metro area a city fell within, or the state. Related tables like state_pop and personal_income stem from considering any questions a user might have about why certain trees characterize a city, the condition of those trees, and whether the trees are introduced/non-native or naturally occurring.

Sample Data

We populated our database with data collated from several sources. The largest source of data was an inventory of 5 million city trees across 63 major U.S. cities from the Dryad Data Platform. We limited the cities considered to 15 out of 63 to create a more manageable dataset, as well as limit selection to cities with the most complete data. The information we focused on pulling from this dataset included tree common names, scientific names (i.e. combined Genus and species classification), state, city, coordinates, native status, physical condition of each observed tree, height in meters, and diameter at breast height in centimeters. Information pulled from U.S. Census data included total state population, foreign born population, and the percentage of total population that is foreign born. The BEA dataset provided us with information about personal per capita income for 2021 based on greater metropolitan areas; we manually categorized each city from the Dryad dataset into their corresponding metro areas. Information pulled from the CDC dataset focused on the prevalence of chronic health issues by city such as mental health, high blood pressure, and diabetes. Pictured below is a sample result set from the tree_measurements table:

Figure 2. *Tree Measurements Table Screenshot*

city_id	longitude	latitude	tree_id	dbh_cm	height_m	condition_id	native_id
803	-76.70799430	39.36215479	21	125.73	278.87	2	2
803	-76.70814780	39.36216514	32	65.53	262.47	3	2
803	-76.70642810	39.36231399	21	91.95	196.85	3	2
803	-76.70781200	39.36216594	32	82.80	246.06	1	2

Views / Queries

Query Name *(# Required)*	JOIN *(4)*	FILTER *(3)*	AGGREGATE *(2)*	LINKING *(1)*	SUBQUERY *(1)*
person_income_and _health	**X**	**X**			
location_population_physical_health	**X**				
avg_tree_height	**X**		**X**		
most_mature_tree_by_city	**X**		**X**		
dead_poor_condition_trees	**X**	**X**		**X**	**X**
introduced_trees_immigrant_pop	**X**	**X**	**X**	**X**	**X**
emerald_ash_borer_tracker	**X**	**X**			**X**
Total	**7**	**4**	**3**	**2**	**3**

Changes from Original Design

The first major change was our decision to use 15 cities instead of the original 10 selected so that each entity in our database would have at least 15 records. A few smaller changes included using more recent personal income data from 2021 instead of 2020, and removing the year of most recent observation for an inventoried tree. The last notable change was the structure of our ERD from proposal to final design. We began with 10 tables–specifically separating tree geolocation data (i.e. latitude and longitude), address for observed trees, immigrant population, and population health outcomes into their own tables. It wasn't necessary to have an address if coordinates were provided, so coordinates were combined into the tree_measurements table because they depended on both a city_id and tree_id. The health outcomes data was dependent on both city_id and state_id, so its data was merged into the city_states linking table. Further restructuring merged the immigrant_population table data into the states table–now state_pop. Our final design has 9 tables that are less atomized, but easier to navigate as a result.

Database Ethics Considerations

We do not have significant data privacy or ethical concerns about our database, but there are some important things to keep in mind. For example, one privacy concern is that some of the trees that are included in our database are precisely geolocated in residential areas. It's important to know that not all of the trees will be on public property and that in any of the data collection

projects, one must be conscious of requesting permission and access to trees in order to collect data. Additionally, although census data is the most comprehensive form of population, income, and race data in the U.S., it is known for erasing marginalized groups and potentially resulting in their being underserved. Our database might perpetuate that bias as a result. Using the census data gives us a decent estimate but is not a full and accurate representation. We are taking all of these things into consideration and will make it transparent when publishing our database.

Lessons Learned

Communication is a must. Having weekly meetings and regular goal check-ins is something our group could have improved upon throughout the semester. Making time to meet outside of the discussion section was tough. We also think there could have been a better way to collaborate on the ERD and queries. There was limited collaboration when it came to actually creating the ERD and queries. It was more convenient for a few to create the database, but this caused a bottleneck in the workflow since the rest of the team would have to wait until database creation to start their queries.

However, there were also many things our team did well that we would recommend to future teams. We chose to base our project off pre-existing data on tree types, sizes and conditions in the biggest cities in America. Using pre-existing data enabled us to understand how to make real data work in database design. Additionally, our preliminary ideas were more realistic than if we hadn't done so. Getting a head start on our ERD with the Project Proposal also helped us out in the long term. By envisioning our data dn our normalization process early on, it forced us to think about our project long-term, even in the early stages. This eventually benefitted us on the planning side and helped us create a cohesive project from the start.

Potential Future Work

If we were to continue working on this database, we think a much broader cross-section of cities inventoried would improve results. Only pulling 25 entries each from 15 city tree inventories out of 63, significantly limited the diversity of trees in our database. We had about 134 different trees or plant classifications, but more would offer insight into what cities and states consider environmentally important. For example, tree data from Honolulu inventoried snag trees (considered useless dead wood to some, but of importance to nesting animals), which speaks to the goals of tree management changing based on the city.

References

Centers for Disease Control and Prevention. (2021). 500 cities: City-level data (GIS friendly format), 2019 release. Centers for Disease Control and Prevention. Retrieved October 13, 2022, from https://chronicdata.cdc.gov/500-Cities-Places/500-Cities-City-level-Data-GIS-Friendly-Format-201/dxpw-cm5u

McCoy, D., Goulet-Scott, B., Meng, W., Atahan, B., Kiros, H., Nishino, M., & Kartesz, J. (2022, August 31). A dataset of 5 million city trees from 63 US cities: Species, location, nativity status, health, and more. Dryad Data -- A dataset of 5 million city trees from 63 US cities: species, location, nativity status, health, and more. Retrieved October 10, 2022, from https://datadryad.org/stash/dataset/doi:10.5061/dryad.2jm63xsrf

U.S. Bureau of Economic Analysis [BEA]. (2021). Personal income by county, Metro, and other areas. Retrieved October 10, 2022, from https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas

U.S. Census Bureau. (2021). Census Bureau Tables. Explore census data. Retrieved October 13, 2022, from https://data.census.gov/cedsci/table?q=immigrant+population&g=0100000US%240400000