Project Report

INST327

Professor Pamela Duffy

By: Elenna Mach, Mohammad Shahid, Justin Xu, Priscilla Kreger

Introduction

When creating our database, we wanted to ensure that finding tickets and violations is clear and concise in an overall condensed way in our database for the most efficiency. We structure our database around individual traffic tickets. Each ticket represents an instance of a traffic violation and includes relevant information such as violation type, location of violation, officer code, and fine details. Notices are sent to the recipients of traffic tickets and can include multiple tickets. Through this design, we can search for traffic violations by any parameters included in a traffic ticket. Our database is derived from our primary data source the ProPublica Chicago Traffic Violations Datastore. This dataset contains records from 2007 to the current day. To limit the scope of our project we will only include data from within the past 6 months which ensures our data isn't superfluous and most useful for users looking for tickets and violations recently after they think they received it. Our system should enable cross-location studies of traffic patterns in Chicago which can promote the enactment of better traffic policies for the city. Areas with high concentrations of citations can be studied to understand pain points in the city. Increased transparency between the traffic authorities and citizens promotes trust and encourages safer driving. It is our team's vision that the data our system provides will be used to enrich the lives of Chicago residents and improve driving conditions for commuters.

Database Description

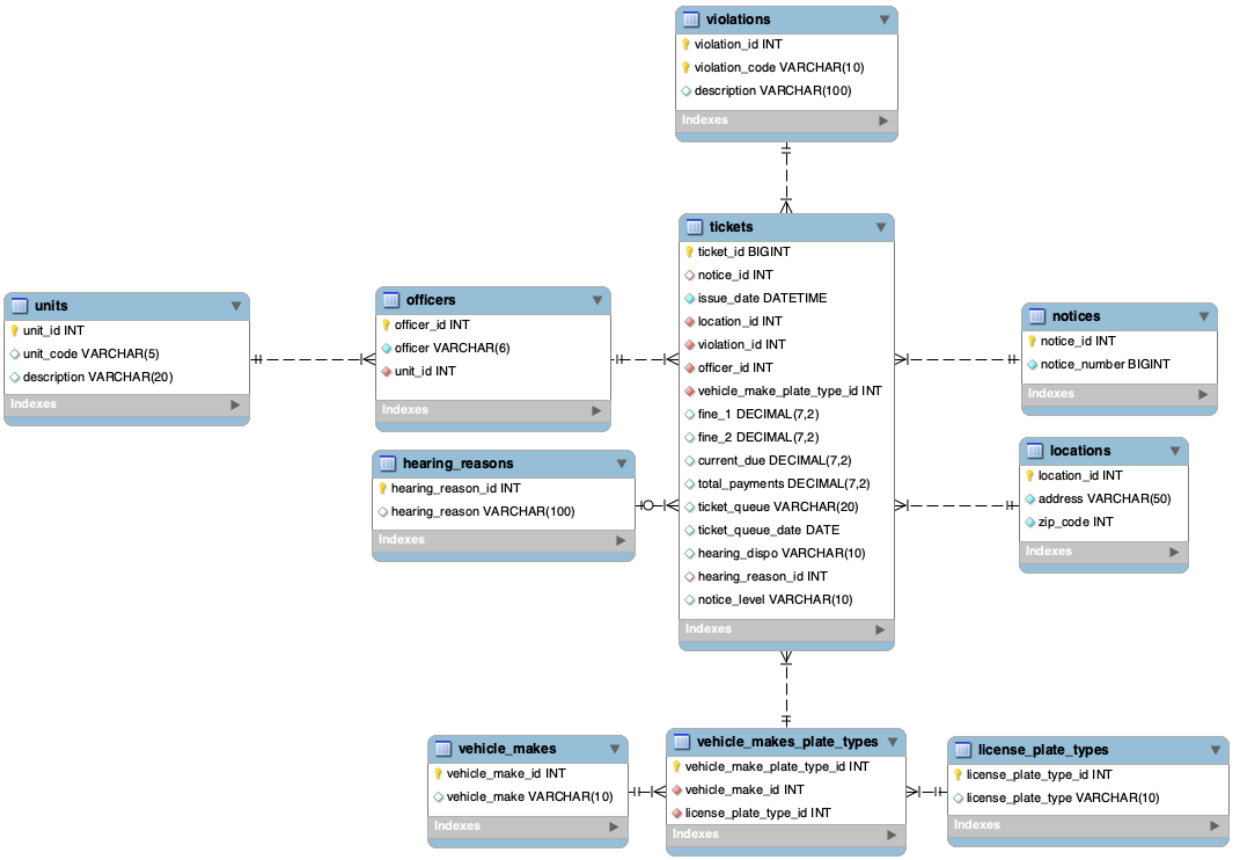
Physical Database

Our physical database backup is included in our submission as "Normalized DB Backup.sql." It includes all sample data and views.

Sample Data

Our Initial dataset contained 1993 rows of citation data. Our team filtered the data down to 152 rows by keeping only traffic tickets with an amount paid of zero, an amount due of zero, and a valid hearing disposition. This quantity of data allows us to populate our database with enough sample data to test our design without overwhelming reviewers and testers. In our submission, we have included the original unfiltered dataset and the cleaned dataset. Our "hearing_reasons" tables are under the required 15 rows but that is due to the limitations of the full dataset that we were provided with. In a live scenario, this would not be an issue. Thus our team has deemed it important to leave in.

Logical Design



Views/Queries (CRUD)

View Name	Req A (4)	Req B (3)	Req C (2)	Req D (1)	Req E (1)
chicago_norm_all	X			X	
officers_units_all	X				
officers_per_unit	X				
vehicle_makes_plate_types_all	X			X	
tickets_per_notice	X	X	X		
officers_per_unit	X	X	X		
tickets_from_small_units	X	X	X		X

Changes from the Original Design

In our original design when we first worked with this data, we sought to analyze parking violations, speeding citations, and running red light violations in the City of Chicago to draw conclusions regarding the relationship between socioeconomic status and traffic citations. We then went on to generalize this goal by instead aiming to use our database to be able to contribute towards more effective traffic policies by analyzing our data.

In terms of normalization, we significantly reduced our number of entities to 7, including violations, tickets, officers, units, hearing reasons, notices, locations, license plate types, vehicle makes, plate types, and vehicle makes. Our final database excluded queues, fines, payments, dues, and zipcodes which were unnecessary tables in our initial database. Our initial diagram had many attributes such as tickets_1 and tickets_2 as well as current_date, total_payments, etc. that we consolidated to make our final table simpler and easier to navigate. We added a new table vehichle_makes_plate_type which contains vehicle make and license plate type to also further consulate our data. Column names were then updated to be simpler and more accurate.

Reorganizing was a key change we focused on from our initial database. We changed the relationships to simplify and consolidate our data. Tickets now store more attributes directly which helps users navigate our database. The final database was also denormalized so previous attributes such as queues, fines, and payments, ended up falling into the tickets table.

Database Ethics Considerations

Throughout our project, we ensured that data privacy was preserved and prioritized by excluding personally identifiable information, which minimizes major privacy risks. The Chicago traffic citation database contains personal data such as license plate numbers. Even though this is public data, it is still personal and identifiable information and it was not necessary in our database for our purposes so it was most ideal to leave it out. This is another value we found to be important to us and all had the same drive to reinforce data privacy further. In terms of fair use, the project remains compliant by limiting the dataset to internal analysis and ensuring that no data is shared for commercial purposes without proper permissions. These steps ensure that the database is not only a comprehensive analytical tool but also a model for ethical and responsible data management. We have a shared motivation of striving to change the trend of citation databases to be less redundant and account for the diversity and equity flaws we have studied. This has led us to work hard on a database that not only alleviates these issues but is built on the ethical and functional values each team member holds.

Lessons Learned

We learned the importance of normalization as it is a key aspect to reducing redundancy and having maximum efficiency since our database was initially over-normalized and not all data points had matching entity or attribute names in our ERD. This was an aspect we had the most trouble with and needed to seek TA and professor help and relentlessly revise to keep our data simple and practical for future users looking for traffic citations. As we learned that over-normalization might complicate queries, while under-normalization can lead to data

redundancy, our ERD had to be edited many times. Furthermore, defining the primary and foreign keys while ensuring we had a reasonable amount of keys, overall drove us to create a simple yet detailed database. Understanding relationships and their constraints within our code also proved to be a meticulous and detailed task. We learned, however, that lack of proper definition and consensus on these key aspects can lead to both an over-complicated code and database. Moving forward we know thorough consideration and planning, along with many revisions, and consistency, makes the best database for users.

Potential Future Work

Our database increases the transparency in the traffic citation system and enables the development of more effective traffic policies by studying citation data. We found that traffic citation systems in many states overall were redundant and inefficient. One significant concern of the existing Chicago traffic citation database is the placement of traffic cameras, which could unintentionally reflect biases. One of our team's core values is inclusion and equity which is a key motivator and value of creating our database. The design of the Chicago traffic citations database is inherently objective due to its reliance on automated data collection, such as red-light and speed cameras. However, achieving true inclusiveness requires reassessing potential biases in the dataset. For example, cameras may be disproportionately located in areas with lower socioeconomic status or higher minority populations, leading to over-monitoring of these communities. Including geographic and demographic data in the database enables the identification of such patterns, supporting a more diverse and equitable analysis. Additionally, it is crucial to ensure the dataset represents all neighborhoods and zip codes, including those with lower citation density, to avoid skewed results. Examining citation trends across neighborhoods, zip codes, and types of violations over time helps ensure the database reflects the full social and economic diversity of Chicago. Additionally, it may be interesting to analyze if there is a significant number of non-Chicago residents who are committing these traffic violations. To fulfill this hypothesis, our database and further research would have to include license plates to analyze the states they are from. Further, this future research would have to consider data privacy again, since license plates are personally identifiable pieces of data.

Works Cited

ProPublica. (n.d.). ProPublica Data Store | City of Chicago camera tickets and warnings….Propublica Data Store. https://www.propublica.org/datastore/dataset/city-of-chicago-camera-tickets-and-warnings-data