INST327-0101 Database Design and Modeling

Project Final Submission

Team: 0101-G1

Members: Eknoor Kaur, James Droter, Lakshya SK, Jordon Farrell

**Introduction**

Our project is focused on analyzing Chicago traffic citations to gain insights into traffic violations, patterns, and trends within the city. The database will consist of a wide range of attributes related to traffic citations, including but not limited to, citation number, date and time of the violation, location, vehicle make and model, violation type (e.g., speeding, red light violation), driver demographics (if available), and officer information. We aim to cover several years of data to capture longitudinal trends and seasonal variations in traffic violations.

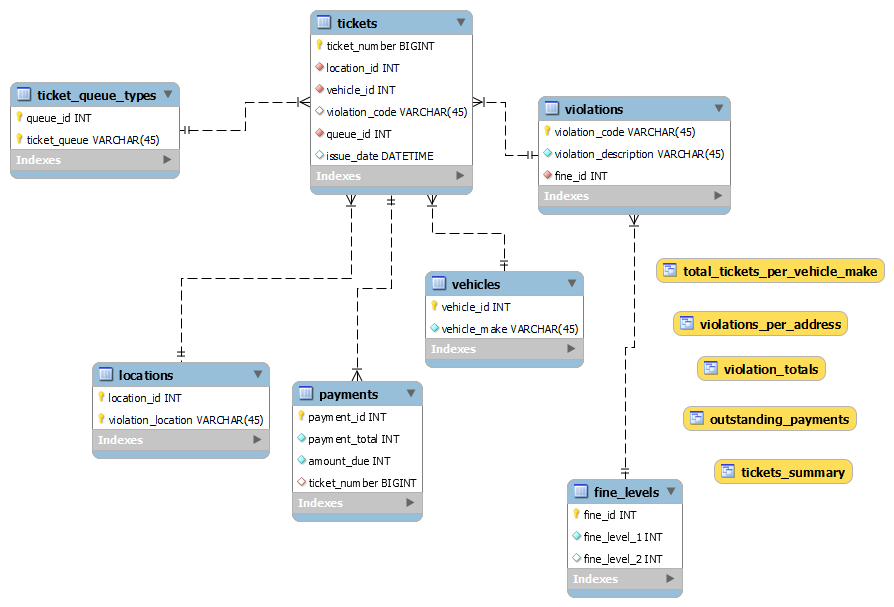
The scope of our research involves compiling and organizing data from many sources, such as public records, Chicago Police Department databases, and, if available, data from traffic cameras and sensors. We hope to utilize this data to identify areas that are hotspots for specific violations, look at demographic patterns in citations, and assess the efficacy of traffic enforcement tactics.

We chose this topic because traffic citations provide valuable insights into urban transportation dynamics and safety issues. Understanding traffic violation patterns can inform policy-making, resource allocation for law enforcement, and urban planning initiatives aimed at improving traffic safety and mobility in cities like Chicago.

**Logical Design**

When designing our database, we identified seven entities: tickets, ticket queue types, violations, locations, payments, vehicles, and fine levels. Each of the entities has the following attributes:

* Tickets
  + Ticket number (PK), location ID (FK), vehicle ID (FK), vehicle ID (FK), Violation code (FK), queue ID (FK), Issue date
* Ticket queue types
  + Queue ID (PK), Ticket queue
* Violations
  + Violation Code (PK), Violation description, Fine id
* Vehicles
  + Vehicle ID (PK), vehicle make
* Locations
  + Location ID (PK), violation location
* Payments
  + Payment id (PK), payment total, amount due, ticket number (FK)
* Fine levels
  + Fine id (PK), fine level 1, fine level 2



The relationships in our logical design are as follows:

1. Tickets one-to-many Payments
2. Violation one-to-many Tickets
3. Vehicle one-to-many Tickets
4. Location one-to-many Tickets
5. Ticket Queue one-to-many Tickets
6. Fine levels one-to-many violations

We centered the design around the tickets table because the ticket is the main piece of information the database is tracking. From there, we split the information contained in the tickets into tables based on their purpose and relationship with the ticket.

**Physical Design section**

When creating the physical database, we designed it around the tickets table since the ticket contains all the information in the database. We split the information on the ticket into separate tables based on the first, second, and third normal forms, making sure to adhere to those rules. After creating and refining the ERD, we imported the data into the database by first importing the entire dataset as its own table. From there, we used inserts to take the data from the main dataset and place it in its appropriate table. After validating that the data was imported correctly and that it meets the project requirements, we created the select queries as views. Finally, after completing the import and views, we created a full backup of the database using the data export tool built into MySql Workbench. The last step for creating the database was to make sure that the backup script would run without issues on multiple computers. After running the backup script on multiple computers and verifying that the data was restored correctly, the database was completed.

**Sample Data**

To filter out rows from the original dataset and create a usable sample for import into the MySQL database, we employed several strategies as outlined:

* Sampling: Instead of analyzing the entire dataset, we opted to randomly sample a subset of the data. This approach allows us to work with a manageable portion of the data while still retaining its overall characteristics. For this task, we aimed to include roughly 50 rows from the original dataset.
* Filtering: We filtered the dataset based on relevant criteria to focus on specific subsets of data. This ensures that the data we import into the database is relevant to our analysis objectives.
* Duplication Removal: We took measures to remove duplicates, handle missing values, and standardize formats. This step ensures that the dataset is clean and free from unnecessary data volume, enhancing the efficiency of our analysis process.

For our sample dataset, we selected rows that meet the criteria of our analysis requirements. We ensured that this sample contains enough rows to fulfill the row requirements for certain tables in our database, such as 15 rows for non-linking tables and 30 rows for linking tables. This ensures that our database has sufficient data for a meaningful analysis.

Columns not included in the database:

* Unit
* Office
* Hearing Dispo (Disposition) and Hearing Reason
* License Plate Type

**CRUD (Views/Queries)**

These queries/views are designed to extract and summarize specific information from the database regarding tickets, violations, payments, and vehicle information. The query description is as follows:

* total\_tickets\_per\_vehicle\_make: This view calculates the total number of tickets per vehicle make, and only includes vehicles make with more than 5 tickets. It joins the `vehicles` and `tickets` tables on the `vehicle\_id` column, groups the data by `vehicle\_make`, and counts the occurrences of each make.
* violation\_totals: This view summarizes violation data by calculating the total quantity of each violation description and the total fines associated with each violation. It joins the `tickets`, `payments`, and `violations` tables to gather information on fines and violation descriptions. The fines are summed up, and the results are formatted to display the total fines in a currency format. It also filters out violations with zero fines and orders the results by violation quantity in descending order.
* tickets\_summary: This view provides a summary of tickets, including the ticket number, violation description, and ticket status. It joins the `tickets`, `violations`, and `ticket\_queue\_types` tables to gather information about each ticket's violation and status. It excludes tickets with a status of "Dismissed" or "Warning" and orders the results by ticket number.
* outstanding\_payments: This view lists tickets with outstanding payments, showing the ticket number and the amount due. It joins the `tickets` and `payments` tables to identify tickets with unpaid amounts and orders the results by the amount due in descending order.
* violations\_per\_address: This view calculates the total number of violations per violation location, but only includes locations with more than 2 violations. It joins the `tickets`, `locations`, and `violations` tables to gather information about violation locations and counts the distinct occurrences of each location. It orders the results by the total number of violations in descending order.

| View Name | Req. A | Req. B | Req. C | Req. D | Req. E |
| --- | --- | --- | --- | --- | --- |
| total\_tickets \_per\_vehicle\_make | x | x | x |  |  |
| violation\_totals | x | x | x | x | x |
| tickets\_summary | x | x |  |  |  |
| outstanding\_payments | x | x |  |  |  |
| violations\_per\_address | x | x | x |  |  |

**Database Ethics Considerations**

Ethical considerations surrounding diversity, equity, inclusion, data privacy, fair use, and other aspects influenced our project analyzing Chicago traffic citations immensely. We made sure to craft our database architecture and sample data plan to ensure the representation of diverse areas and populations, prioritizing equity and fairness. We emphasized addressing potential biases in enforcement tactics, safeguarding personal identification information, and remaining transparent about our methods and findings. These considerations guided our project, shaping our database design, sample data selection, and engagement with stakeholders to conduct our analysis responsibly and ethically, contributing to a more just and equitable understanding of traffic violations in Chicago.

**Changes from the Original Design**

Throughout the project, we incorporated feedback from the TA's to improve our work. Initially, we removed two tables - ticket violations and ticket locations - as they did not adhere to the three normal forms. Additionally, we developed a new table to house fine levels and linked it to the violations table, establishing a one-to-many relationship. Our final modification was to eliminate the zipcode from the database, as it referred to the registration zipcode of the cited vehicle and did not align with the addresses in the locations table.

**Lessons Learned**

Throughout the project, the team gained valuable experience when it comes to collaboration and technical skills development. Working together helped learn the importance of clear communication, effective time management, and the ability to integrate different schedules. Through regular coordination of our efforts and planning, we ensured all team members were up-to-date on the task at hand and everyone was able to contribute effectively. This project also provided us with a foundation in SQL. Through the semester, the team skills from querying to database design. The incorporation of feedback from mentors also played an important role in our approach through the project. Project Status meetings afforded the opportunity for us as a team to improve the database. The lessons learned were the improvement on collaboration skills, SQL technical skills, and time management usage.

**Potential Future Work**

The addition of future database entities would prove advantageous for Chicago. The introduction of driver demographics, traffic camera data, and weather conditions into our database will improve functionality and impact. Driver demographics allows for targeted law enforcement and education to different groups. Camera data would provide real-time insights which would help with overall response time for an issue. Weather conditions helps to analyze the reasoning of the violations, evaluating another potential variable. The improved database would be a powerful tool for law enforcement in the city of Chicago.

**Sources:**

City of Chicago camera tickets and warnings data. (2022). [Dataset]. In *ProPublica Data Store*. <https://www.propublica.org/datastore/dataset/city-of-chicago-camera-tickets-and-warnings-data>