INST 327 - Final Project Submission
Team 4: Project Final Report
December 14, 2023
Jared Sherman, Rileigh Thorward, Jason Candila, Giulia Hoorens van Heyningen, Tiara Eltrevoog

Introduction

Car crashes have been a constant source of fear, tragedy, and death for a very long time. Millions of Americans get in a car every day, whether they are the driver or the passenger. We use cars to go to work, to get home, to go to events, and anything and everything in between. Driving is an integral part of many lives and one that has drastic implications if something goes wrong. We wanted to create a car accident database that could store information about collisions to see if we can find a deeper meaning. We selected this topic because it addresses an incredibly important aspect of public safety, it has the potential to reduce injuries, and save lives. Our goal was to understand the factors that contribute to accidents, the types of vehicles involved, and the conditions in which accidents occur. As public transit isn't a viable option for millions of Americans, their safety on a daily commute should be an absolute priority. Furthermore, understanding which models of vehicles are commonly in dangerous accidents can help citizens make more informed decisions on their choice of car. Many citizens don't have the full picture of a car's reputation when car shopping, and the companies are not always entirely forward with the customers. Moreover, this database can also help the automotive industry implement new technology for car safety. As we progress as a society and further move through this technological era, implementing safety technology is essential when manufacturing cars. In fact, safety should be the number one priority for all car manufacturers, and our data will help show which companies need to invest more time and money into just that. Ultimately, our car accident database will serve as a valuable tool in promoting road safety and improving the overall well-being of individuals and communities.

Database Description

The database collects information based on accidents happening within the state of New York, including what we think to be crucial details when analyzing the accidents that have occurred by showing data regarding pre-crash positioning, the time of the accident, the type of vehicle involved, etc… The database includes 10 tables and 613 cases each with their unique data organized by IDs and descriptions.

Logical Design

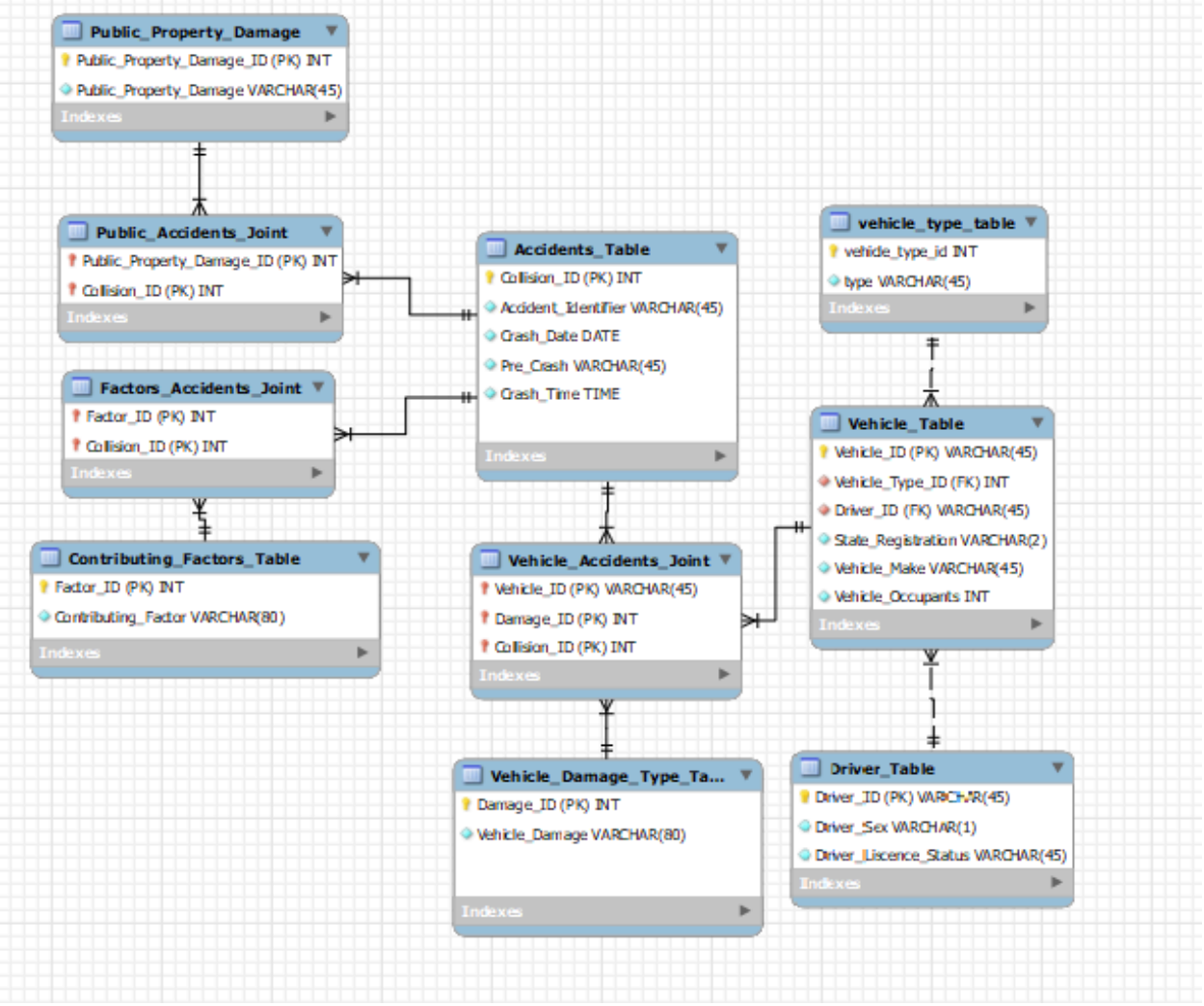

Image 1: Entity Relationship Diagram for collision data.

The purpose of the ERD was to help us visualize the final table split into different parts to better make sense of the original dataset we were tasked with using. Due to this being an important blueprint for future ideas, we were focused on making a clear diagram with simple yet robust concepts to enable easy understanding while accounting for all possible scenarios.

The dataset consists of different types of tables. Tables such as accidents and vehicles include crucial information about all 613 cases. The design of these tables differs, such as how the vehicle table depends on two smaller tables which usually contain details about another important section of information. For example, the vehicles table relies on the driver table which contains its own information. We agreed to this design to reduce the amount of information that would show up on one table at a time, we were able to reduce the amount of max columns to five, ideally helping the user take in the kind of information that there is. Joint tables were utilized to create a bridge to the smaller detail tables and the larger tables to highlight a many-to-many relationship. This was done because of the amount of repeating values that would occur throughout the table. For instance, in nearly a hundred cases, there was an accident where a factor of “driver

inattention" was involved, we were able to attach an ID to this description in the contributing factors table and use that as a foreign key to a join table attaching it to the case involving that information.

The relationships established reflect our want for an efficient and clear database, being able to logically associate descriptions with the corresponding case utilizing the skills we have learned from the class and our planning done in the form of normalization.

Physical Database

Our database seeks to provide a summary of several accidents that have occurred in New York State. To do this we have provided relevant and non-redundant columns of data which includes details on these accidents varying from the kind of vehicle involved to the believed factor that caused it. The kind of information we included was relevant to the audience we designed it for, that being governments, insurance companies, and even drivers. These kinds of audiences will utilize this information in different ways, so we were required to create a robust database. With governments in mind, we included tables like public_property_damage_tables, vehicle_type_tables. For insurance companies, we wanted to include the driver_information_tables. Finally for driver education, we included the contributing_factors_tables. These choices reflect our want to create a database for the purpose for education and informational purposes.

Sample Data

Our data is derived from the 'Motor Vehicle Collisions' Dataset, which originally contained over 3 million rows. Using the Pandas module within Python, our first step was to remove the attributes that we specified would be irrelevant to the scope of our project, including: 'TRAVEL_DIRECTION', 'DRIVER_LICENSE_JURISDICTION', 'VEHICLE_YEAR', 'VEHICLE_MODEL', and 'PUBLIC_PROPERTY_DAMAGE_TYPE'. We managed to refine the data to 500 rows of unique collisions. Our first step was to filter by state and date. Our final dataset only includes collisions that occurred between 1/01/2021 and 12/31/2021, where the 'STATE_REGISTRATION' was New York. We also removed rows containing null or insignificant (such as '0' or 'No Damage') values in the following attributes: 'CONTRIBUTING_FACTOR_1', 'DRIVER_SEX', 'DRIVER_LICENSE_STATUS', 'PRE_CRASH', 'VEHICLE_MAKE', 'VEHICLE_TYPE', 'VEHICLE_OCCUPANTS', 'VEHICLE_DAMAGE', 'VEHICLE_DAMAGE_1', 'VEHICLE_DAMAGE_2', 'VEHICLE_DAMAGE_3'. This allowed us to narrow the scope of our data to more significant collisions, while also ensuring that the data we did include was complete. We consider significant collisions to be those which included passengers and inflicted vehicle damages. However, despite significant data cleaning, the dataset was still far too large. Thus, we removed alternating rows until we were left with 612 unique collisions, then dropped the last 112.

Collision_ID (PK)	Accident_Identif...	Crash_Date	Pre_Crash	Crash_Time
1	4380859	2021-01-01	Going Straight Ahead	16:25:00
2	4381324	2021-01-01	Going Straight Ahead	02:58:00
3	4380871	2021-01-01	Going Straight Ahead	23:35:00
4	4381072	2021-01-02	Going Straight Ahead	19:18:00
5	4381396	2021-01-03	Going Straight Ahead	00:20:00
6	4381252	2021-01-03	Going Straight Ahead	12:00:00
7	4381812	2021-01-04	Going Straight Ahead	08:00:00
8	4381618	2021-01-04	Stopped in Traffic	11:30:00
9	4381828	2021-01-04	Going Straight Ahead	20:31:00
10	4381730	2021-01-05	Going Straight Ahead	13:27:00
11	4383778	2021-01-06	Passing	08:03:00
12	4382025	2021-01-06	Slowing or Stopping	17:20:00
13	4382882	2021-01-07	Making Left Turn	19:48:00
14	4382494	2021-01-07	Slowing or Stopping	11:33:00
15	4382554	2021-01-08	Going Straight Ahead	11:40:00
16	4382585	2021-01-08	Going Straight Ahead	16:16:00
17	4382900	2021-01-09	Going Straight Ahead	18:40:00
18	4382795	2021-01-09	Going Straight Ahead	14:32:00
19	4382916	2021-01-10	Making U Turn	22:45:00
20	4383511	2021-01-11	Going Straight Ahead	05:20:00

Image_2: A snippet of the accidents table displaying various accidents in the database.

Views / Queries

View Name	Join	Filter	Aggregate	Linking	Subquery
accident_count_by_date		X	X		
accidents_info	X	X		X	
damaged_vehicles	X	X	X	X	
high_risk_drivers	X	X		X	X
public_property_damage_info	X			X	

Query 1: In this query, I created a view that aggregates and filters the data so that we can count the number of accidents for each of the crash dates. I excluded dates that only had one accident.

Query 2: This created view joins the accident_table with the factors_accidents_joint and the contributing_factors_table to give information about the accidents, like collision ID, crash date, crash time, and contributing factors.

Query 3: This view aggregates and filters data in order to count the number of damaged vehicles for the vehicle make. I only focused on those damage types that were categorized as front damage. The vehicle_table, vehicle_accidents_joint and the vehicle_damage_type_table were all joined in this query.

Query 4: This query creates a view that determines what high-risk drivers were involved in accidents by including the driver details, vehicle details, and accident details. I focused on only including pre-crash results that were defined as ‘changing lanes’. The driver_table, vehicle_table, vehicle_accidents_joint, and accidents_table were all joined in this query.

Query 5: This view gives information about the public property damage and uses the public property damage ID, collision ID, and type of damage, by joining the public_property_damage table with the public_accidents_joint and accidents_table.

Changes from the original design:

Our initial idea has stayed fairly consistent throughout this entire process. In terms of scope and focus, we’ve maintained the same entities, such as Accidents, Public Property Damage, Drivers, and Vehicles. The exclusion of certain entities – travel_direction, driver_license_jurisdiction, and vehicle_year – is also consistent with the initial proposal. We have done some data reduction and filtering to address the feedback we received from our Project Proposal and Project Progress Report. After table normalization, we of course had to include the creation of joint tables since there were multiple many-to-many relationships we identified. When creating relationships we also had designed the vehicle damage types table to be connected to the vehicle table, however, due to feedback, we changed it to connect it to the accidents and vehicle joint table because we originally did not consider the many-to-many relationship these two tables had. A small change during this time included making sure all values were non-null. During the data import, certain primary keys that were in the original table were not able to be used as primary keys due to there being duplicates. The accident table’s “accident_identifier” was perceived to be a primary key, but due to the oversight, a new primary key of simple int incrementing values was needed and therefore several joint tables were changed.

Diversity, Equity, and Inclusion Considerations:

When developing our database, we prioritized Diversity, Equity, and Inclusion (DEI) considerations to create a comprehensive and unbiased representation of driver profiles. Recognizing the significance of demographic attributes, such as "Driver_Sex" and "Driver_License_Status," we intentionally incorporated these variables to ensure a better understanding of diverse driver backgrounds. This strategic inclusion aligns with ethical standards, aiming to avoid biases and promote inclusivity in our dataset. By incorporating DEI principles, our database not only adheres to ethical guidelines but also strives to deliver insights that reflect the varied characteristics of the population under study. We also incorporated the contributing factors data into our final product to prevent any possible judgment or discrimination from happening that could be gathered from our sex and status demographics.

Data Privacy, Fair Use, Other Ethical Considerations:

A major focus of our project was making sure to consider data privacy, ensure fair use, and adhere to other ethical considerations. The nature of attributes like "Driver_Sex,” “Vehicle_ID”, and "Driver_License_Status," suggested efforts to protect user information. We made sure to focus our results and details of the accidents to other attributes like property damage and contributing factors to protect users information and privacy. In terms of ethical/legal

considerations, private entities like law enforcement would need to grant permission to someone in order to get the dataset, especially any data with private information. However, we would not have to worry about this since it is assumed that privileges and access have already been granted to the college/professors since the CSV files have been available and approved to be downloaded and used by us students.

Lessons Learned:

Throughout this course and working on this project our group has learned many lessons. One of the initial challenges we faced was dealing with a large dataset that contained irrelevant or redundant columns. We learned the importance of thoroughly analyzing the dataset, understanding the domain, and making informed decisions on what data to include or exclude. This process of data cleanup and normalization taught us the significance of a well-structured and efficient database. We also learned the benefits of normalization when creating the ERD and normalization forms. Normalization helped us eliminate data redundancy and create a better database structure. Understanding the normalization process contributed to our ability to design a robust and efficient database. Moreover, working collaboratively throughout this project, taught us the importance of effective communication and teamwork. Discussing ideas, sharing perspectives, and leveraging each team member's strengths was crucial in developing a comprehensive and well-designed database. After each part of the project, we received feedback from instructors, which was an integral part of the process. It taught us to be receptive to constructive criticism, understand the rationale behind suggestions, and implement necessary changes. Adapting our database design based on feedback enhanced the overall quality of our project. Finally, exploring ethical considerations, especially regarding data privacy and inclusivity, highlighted the importance of responsible database development. We realized the significance of incorporating diverse attributes while maintaining user privacy. This experience broadened our understanding of ethical considerations in real world database projects. In specific cases, a lot of lessons were learned from importing the data, as there were a lot of issues we ran into doing this task due to not accounting for these problems.

Potential Future Work:

If we had more time to work on this project we would consider developing a user interface for interacting with the database that could enhance its usability. Creating an application to visualize and query the data would make the database more accessible to users who may not be familiar with SQL. This would be both beneficial for the general public and car manufacturers for additional information on car safety. Additionally, if we had more time we could integrate the database with external data sources or APIs that could provide additional context and enrich the dataset. This could lead to more comprehensive analyses and a deeper understanding of the factors influencing accidents. Another possibility is to develop user training materials and documentation. This would be essential for users interacting with the database, especially for individuals who may not have been involved in its development.