INST327, Section: WB21

Project Report

8/16/2023

Team 4: David Tu, Thompson Nguyen, Joaquim Benitez, Thy Hoang, Edmond Goo

Introduction:

The population of the world has been increasing as every year goes by, and with an increase in population also comes an increase in the amount of vehicles used. From the time span of 1990 to 2022 the U.S. population has increased by 35% and with that the miles traveled by vehicle has also increased by 35% with 278 million registered vehicles and a total of 4.9 trillion miles traveled in the U.S. in 2020 (*Personal Transportation Factsheet*, 2022). With motor vehicle collisions posing a considerable threat to public safety around the world, it is with our project team's experiences of driving that have motivated us to work on a database on the topic of motor vehicle collisions in order to help improve public safety. Being able to identify these trends and patterns with such a large scale dataset will improve public safety not only within New York City but also other urban areas with high density of people and vehicles.

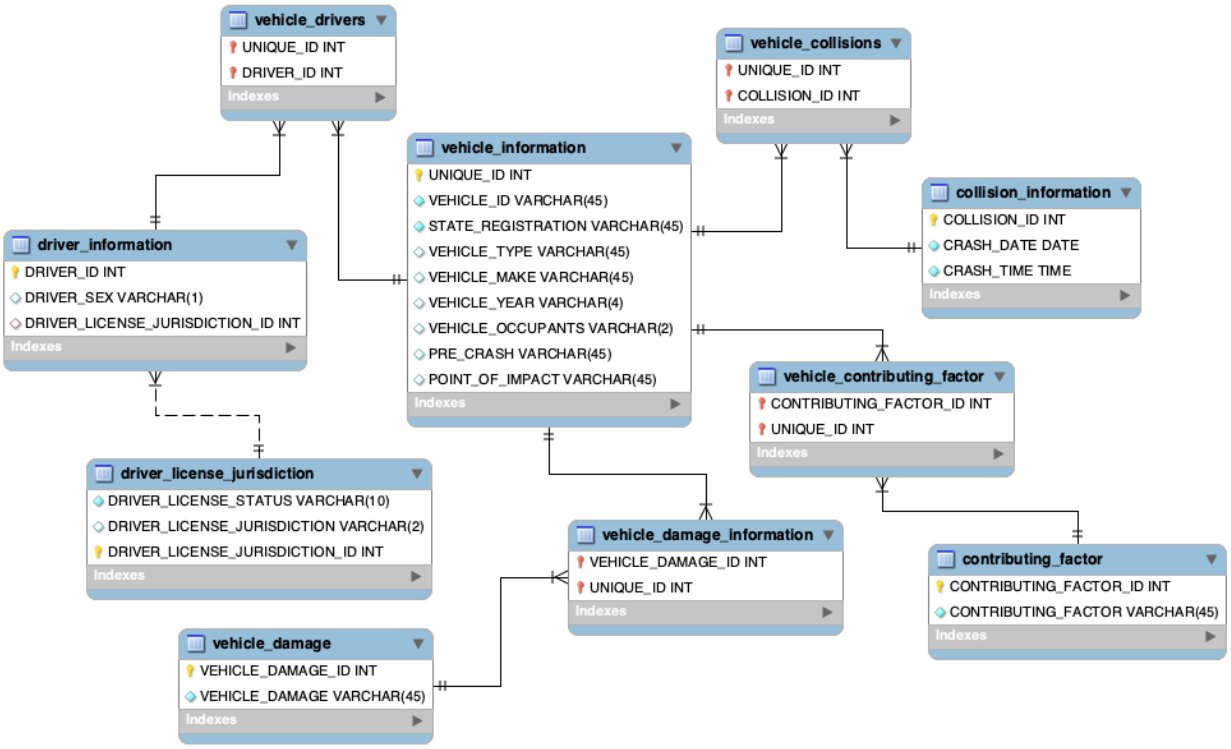
The complexity of motor vehicle collisions demands a well-structured database to effectively analyze and draw insights from the dataset as motor vehicle collisions data can be very volatile. Our project's goal is to provide an optimal dataset that can be used for easier and cleaner data analysis not only for the public but also for various groups such as law enforcement agencies, traffic engineers, urban planners, and even the media to use for a better understanding to be able to address traffic-related challenges. Ultimately, the insights derived from this database can contribute to safer roads and improved traffic management strategies.

Database Description:

The core of our relational database is centered around the "Motor Vehicle Collisions - Crashes" dataset, which contains an extensive amount of information about motor vehicle collisions reported by NYPD since July 1st, 2012. This means that it is necessary to clean up as well as reduce the data for optimal analysis since the original dataset contains such a large amount of data. Our project was able to reduce the original dataset down to the core tables of vehicle information, driver information, collisions information, driver license jurisdiction, vehicle damage, and contributing factors. Along with these tables, we were also able to have

vehicle drivers, vehicle collisions, vehicle damage information, and vehicle contributing factor tables which act as our linking tables for our relational database and entity relationship diagram(ERD) model shown below. Our ERD model is also able to show that we mainly have many-to-many relationships between each core table which are connected through the linking tables listed previously with the exclusion of driver license jurisdiction having a one-to-many relationship with the driver information table. By focusing on these tables and relationships, our database will be able to offer extensive opportunities to be able to efficiently analyze the data to better understand leading causes, high-risk areas, and even peak collision times.

Logical Design:



One of our group's biggest goals was to create a logical design that was extremely efficient and easy to understand for users. With a total of ten tables, we started off with the vehicle_information table which includes the information directly relating to the vehicle and then put the rest of our data through the normalization process in order to generate the tables listed above. For our vehicle_information table, we have listed UNIQUE_ID as our primary key. It is important to note that normally the VEHICLE_ID would also be a primary key to create a composite primary key, however, the data that was contained in the dataset was not stable

enough to be used as a primary/composite key since the data was encrypted by the owners of this dataset which were most likely encrypted to protect privacy.

Physical Design:

Our database illustrates the nature of vehicle collisions and the many different relationships between the driver, vehicle, and collision itself. As such, we've created tables for each of these attributes, but also provided many linking tables to portray those relationships within the database. For example, the driver name, vehicle make and model, and positioning of the vehicle pre-collision are all columns we've created to describe those relational elements. The database relationships will be used to write various queries to explain the questions we've asked in our initial proposal, questions about the nature of the vehicle collisions. The queries will be further explained within the Views/Queries section of the report.

Sample Data:

Our original dataset had over a million entries, so we decided to focus on vehicle collisions that occurred only within the year 2021. We also filtered out any rows with multiple null values so that we get more insight with a more completed entry. Before loading it into the database we rearranged the data so that it would fit into the columns of our ERD and filled in primary key rows. During this process we discovered that some of our tables did not meet the 15 row requirement due to the fact that there were not enough different types of that data in the sample we chose to focus on. For example in our contributing factors table, there were only 12 different types of contributing factors within the rows that we chose to focus on.

Views/Queries:

Query Name	JOIN (x4)	FILTER (x3)	AGGREGATE (x2)	LINKING (x1)	SUB-QUERY (x1)
leading_causes	**X**	**X**	**X**		
make_model_collisions	**X**	**X**	**X**		

collision_time_frame	**X**	**X**	**X**		
collisions_in_seasons	**X**	**X**	**X**	**X**	**X**
older_vs_newer (procedure)		**X**	**X**		**X**
average_occupants_in_collisions		**X**	**X**		
driver_gender_num	**X**	**X**	**X**	**X**	**X**
license_registration_collisions	**X**		**X**		
pre_crash_information	**X**		**X**		
TOTAL:	**7**	**7**	**9**	**2**	**3**

Query 1: Creates a view that displays the amount of collisions for each unique contributing factor sorted by most to least collisions.
Query 2: Creates a view that displays the amount of collisions for each unique vehicle make and its model sorted by most to least collisions.
Query 3: Creates a view that displays the amount of collisions for every two hour intervals of the day sorted by most to least collisions.
Query 4: Creates a view that displays the amount of collisions for every season, ignoring the year.
Query 5: Creates a procedure that lets the user input a year and shows how many collisions occurred before that year and after that year.
Query 6: Creates a view to calculate the average number of occupants in vehicles involved in collisions.
Query 7: Creates a view that displays the number of collisions caused by male drivers compared to female drivers.
Query 8: Creates a view to retrieve driver information for drivers who were involved in collisions and their associated vehicle types.
Query 9: Creates a view to retrieve collision details for collisions involving vehicles of the most common make.

Changes from Original Design:

From our original design, we've changed a number of things to more accurately reflect our dataset. Firstly, we removed the column "vehicle_damage_id" (FK) from the "vehicle information" table because we figured that the relationship between vehicle damages and vehicles is a many to many relationship instead of a many to one relationship, and thus the foreign key was unnecessary. We also removed the "contributing_factor_id" from the "collision_information" table for the same reason. Furthermore, within the "Driver's License Jurisdiction" table, we changed the primary key from "drivers_license_jurisdiction" to making a new PK value name "drivers_license_jurisdiction_ID" since our old PK value was not an INT. Following this, we've decided to also make it possible for the "drivers_license_jurisdiction" column to be null. We've also decided as a team to make all our table names to be lowercase for easier importing and identification. Finally, there is a transitive dependency within the vehicle_information column with the pre-crash and point of impact columns, but after speaking with Professor Duffy, we've decided to not create separate tables for them since we already have enough tables.

Database Ethics Considerations:

As mentioned in our earlier project proposal, we carefully planned out how to set up our database and chose sample data that's fair. We got our information from unbiased sources. We looked at important details and focused on things like where accidents happened and what kinds of cars were involved. The data covers the years 2012 to 2020, which gives us a lot of information to understand motor vehicle collisions better.

Talking about variety, we made sure our data model had a lot of different categories connected to motor vehicle collisions. This included cases where pedestrians, cyclists, trucks, and more were involved. We also considered things like gender and race, which play a big role in how safe traffic is and how accidents happen.

We didn't forget about fairness in our data. We know that different places have different ways of enforcing rules. Some are very strict, while others are more relaxed. When we looked at these differences, we aimed to make sure the data we collected was reliable. With all of these goals in mind, as we got closer to finishing the project, we designed our database to show a complete picture of motor vehicle collisions during that time.

Lessons Learned:

In terms of database design, we learned the importance of normalization to eliminate redundancy and streamline data storage and retrieval. We realized the need to filter and clean the original dataset to focus on relevant information. Choosing appropriate primary keys became apparent as a crucial step to maintaining data integrity and relationships within the database.

We had to adapt and refine our original design to match the characteristics of the dataset, showing the nature of database development. Transitive dependencies within table design also became apparent, leading us to make informed decisions about consolidation.
Creating views and queries to simplify data access and analysis for various purposes became a powerful tool, allowing us to derive insights more effectively.

Creating a large-scale database can help identify trends and patterns that are critical for improving public safety not only in New York City but also in densely populated urban areas around the world.

Potential Future Work:

Any future work we do on this database would depend on our increase in knowledge in databases. I acknowledge that we had to condense the data due to our beginner knowledge of databases and queries and such, or else it would've been impossible to do this project. But, less data means less accurate and more generalized conclusions to answers because we don't have the full picture. In the future, with better knowledge of databases and how to write queries without accidentally deleting a hundred rows, we can use more rows of data to answer our questions and the questions of any others who benefit from this more accurately. We might also be able to write more complex queries to answer more complex questions, questions that the NYPD and the NY government might actually ask. There's also the possibility to increase our scope to more than just New York City. This way, we would be able to help decrease collisions in more than one state.

Work Cited

[Source of database]

Personal Transportation Factsheet. Center for Sustainable Systems. (2022). https://css.umich.edu/publications/factsheets/mobility/personal-transportation-factsheet