INST 327

Project Final

May 15, 2023

Team 0202 - 03: Ken Nguyen, Abrar Hossain, Alfred Mbah, Lexin Deang, Wadi Ahmed

Introduction:

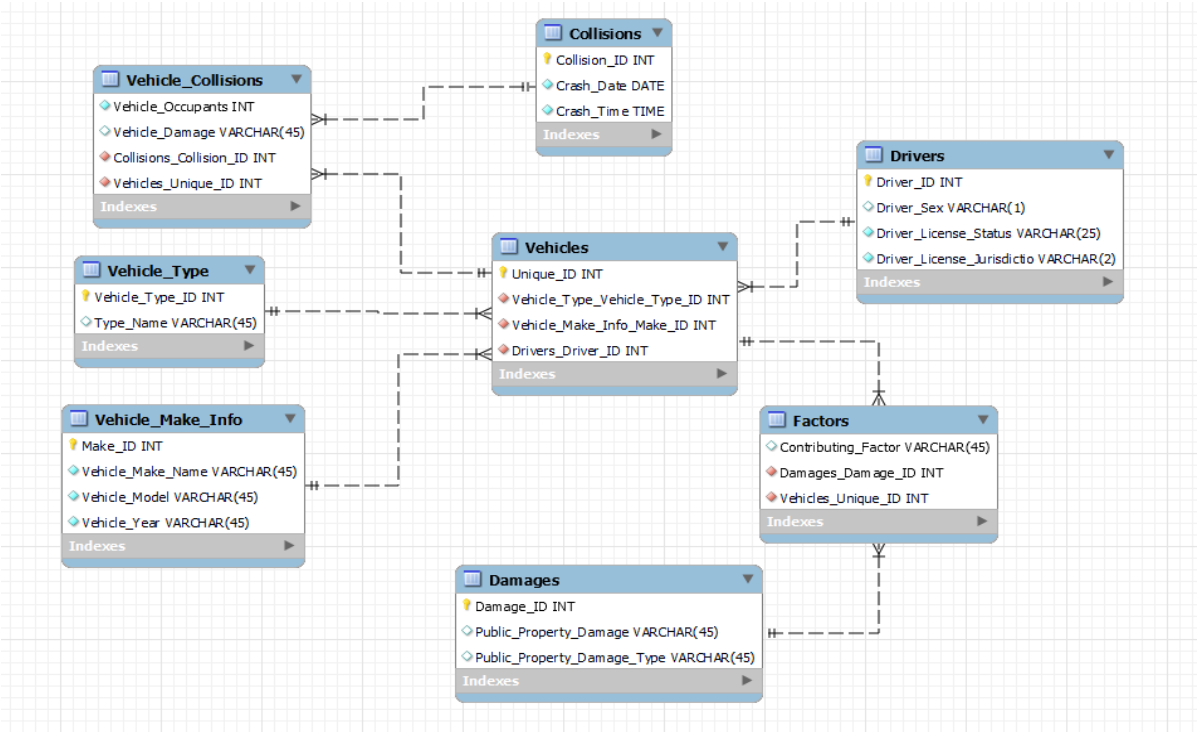
The topic our team chose is motor vehicle collisions. We decided on motor vehicle collisions because we thought it would be the most interesting and everybody in our group shared an interest in cars. The objective we set for ourselves with this database is to improve traffic safety by centralizing all of the retrievable information that is available to us such as traffic-related incidents and motor vehicle crashes. A secondary goal that we also want to aim for is to try and reduce as many vehicle collisions as possible. Doing this would help people in the area be safe while on the road but also reduce the number of reports required for collisions where someone has been injured or killed or where there is $1000 worth of damages.

The database is primarily intended for government entities in New York, such as NYPD and The New York State Association of Traffic Safety Boards (NYSATSB). The NYPD can use this data to put more police coverage in areas with high volumes of motor collisions. This will aid in improving public safety and potentially reducing the impact of future motor collisions. With this dataset, the (NYSATSB) can identify trends in the data and make decisions accordingly to further promote their goal of transportation safety. To summarize, the goal of our database is to improve traffic safety and reduce the number of vehicle collisions that are happening in the area.

Database Description:

Our database will mainly be based on the data collected by the New York City Police Department (NYPD) on motor vehicle collisions. The NYPD is the largest department in the United States with approximately 19,000 civilian employees and 36,000 officers. The data will be coming from New York City, the largest city in the United States with a population of approximately 8.5 million people. The motor vehicle collisions used in our database can be understood in this context as any traffic-related incidents that the NYPD has a report on. The dataset that will come from the NYPD will include all the assumed data relating to traffic incidents with the Motor Vehicle Collision crash tables containing details on the incident event with each row representing an incident event. All of the data that we will be using is going to be tailored to the project and the original database will be used to look for fatal incidents with which the NYPD has information relating to fatal traffic collisions, public safety, and law enforcement to meet our search needs of traffic-related incidents for the Motor Vehicle Collisions. For example, we will be looking at categories such as “VEHICLE_MODEL,” “VEHICLE_TYPE,” “VEHICLE_MAKE,” and “STATE_REGISTRATION” to see how they could be managed in order to improve traffic safety. Since the database has been provided for us, we plan to use it in order to fulfill our search needs. We will be covering over 20 years of banning when looking at this database with the years from 1990 to 2023. This information will help us remove ambiguity, make all our values & elements more defined, and ensure we have an easier time navigating through the database.

Logical Design:



A main objective of our physical design was to make sure it was robust and functional for our users. For the dataset we selected, we opted to make 8 tables with our data. We felt that this was enough tables to get specific information about the data at hand but also broad enough to be able to be used for multiple situations. Our team chose tables that we thought would be most helpful to those who want to see any trends in motor vehicle collisions. Some of these tables include: Vehicle_Type, Vehichle_Make_Info, and Factors. These tables can serve as key identifying factors when looking at why these collisions occurred. Our team tried to provide an easy to read dataset that could hopefully be used to prevent motor vehicle collisions.

Physical Design:

Our database is based on the NYPD database of motor vehicle collisions throughout New York City. It seeks to gain information and data throughout the city to elaborate on the certain motor vehicle collisions that have been occurring throughout. With the development of our physical design, it was imperative that we allow the different tables to be legible and normalized

efficiently in order to captivate a strong sense of analytics with the different views offered. As there was a provided unique identifier from the source database 'UNIQUE_ID', we felt that the primary key 'VEHICLE_ID' was of no use, with the reason that there were data inputs that did not make any sense and correlation with repeating values. Our main table is vehicles, which expands into collisions, factors, types, makes, up to the drivers as well.

Sample Data:

For our data plan, we will be using real data and extracting data from various sources available to us online that include information about motor vehicle collisions in New York City, some of which are directly from the New York City Police Department (NYPD). We will be aiming to utilize the dataset provided to us and any datasets we can extract online in order to assist us in being able to improve driver safety and reduce collisions. Our team also will inquire into other databases for references. After initial research, we have found that there are datasets available from various sites that provide data on motor collisions in the state of New York that can be normalized in order for us to sort out the most common contributing factor for what caused the incident. This would allow us to see potential causes of motor vehicle collisions and then use our dataset in order to provide solutions that would allow us to achieve our database goals of improving driver safety and reducing collisions.

Views and Queries:

Query 1: Create a view that shows the count of collisions per year that pertain to the Toyota make name to determine if interventions throughout these years have been effective or not in the NYC area.

Query 2: Create a view using the database that emphasizes the problems in order to establish ideas to solutions in reducing accidents/collisions.

Query 3: Create a view using the database to pinpoint the different damages done to vehicles that do not include *small* sedans so that comparisons can be made within larger vehicles.

Query 4: Create a view that displays information about the amount of collisions per gender during the evening time.

Query 5: Create a view that displays the counts of damages of vehicular collisions.

Query 6: Creates a view displaying the counts of accidents per month.

Query 7: Since May 1st, 2018, it has been federally required to have a backup camera on all new cars entering the market. This query compares the number of accidents prior to 2018 and after to see the resulting changes from that mandate.

View Name	Req A	Req B	Req C	Req D	Req E
toyota_annual_crash_count	**x**	**x**	**x**	**x**	**x**
factor_count_info	**x**	**x**	**x**		
vehicle_type_and_damage_no_sedan	**x**	**x**	**x**		
gender_collision_evening	**x**	**x**	**x**	**x**	
damage_count			**x**		
pre_2018_rear_crashes	**x**	**x**		**x**	
crash_by_month			**x**		

Changes from Original Design:

There have been many different inputs that were essential to the structure of our overall project. According to the review of our proposal and logical design by our peers, the

UNIQUE_ID was seemingly a confusing aspect of our normalization process. In the perspective of the professor's input of the same instance, UNIQUE_ID was suggested to be implemented as the replacement of VEHICLE_ID to deliver more consistent, meaningful data to the table. On that note, we will change our tables and remove VEHICLE_ID and replace it with UNIQUE_ID within our project's logical design. Another part to be assessed and changed within the project's logical design are various data types that need to be re-established. At a glance, COLLISION_ID should be int, replacing most tiny text into VARCHAR for more leisure with space, as well as the different arrangements of the table to make more sense of normalization and the sophistication of the dataset. From then, we will improve our ERD with changes that assist in relating and corresponding the driver to a one-to-many relationship with vehicles, as drivers can have many different vehicles at hand.

Database Ethics Considerations:

With the proposed database, the team took into consideration the privacy considerations that could occur with some of the information present on the database such as names and street addresses, and by containing that, help reduce data leak risks. As the team states, copyright and fair use doesn't seem to be an issue as the information present is open source and free to use. However, they do include the Vehicle Identification Number, the unique identifier of a vehicle, as a primary key. While this alone does not serve as an identifier, it can be used to trace and locate individuals who are connected to the vehicle. In addition, there doesn't exist a public database of VINs or state licensing numbers, which can aid the teams efforts at getting information about EVs that qualify for the credit. These could lead to vehicle owners' information being leaked and information being taken from them. For a suggestion, the team can use another

identifier for vehicles to uniquely identify them and use in the normalization of the dataset, or to keep a specific pattern.

Lessons Learned:

The biggest lesson that our team learned was how important it is to be able to communicate effectively with each other in order to make our project successful. At the start of our project, there were no clear and defined roles so everybody was lost on which parts were already done and which parts we needed to still complete. As a result of this, many team members were left confused about what their responsibilities for our project were and we had many people working on the same thing. After realizing our problem, we went back and communicated with each other on the roles and responsibilities of each team member which helped a lot. In order to make sure everyone knows what they are doing we created lists and deadlines in the group chat so that every team member can see their roles and responsibilities. Overall, we were able to learn a lot of valuable lessons from working on this project, and am grateful for everything it has taught us.

Potential Future Work:

Our project on motor vehicle collisions can be expanded and improved upon in several ways in the future. One potential idea for the future is developing predictive models; the findings of this research alone can be used as a basis for developing these predictive models that can forecast the likelihood of car crashes in specific areas of the city. Consequently, these models must rely on gathering information on the causes of these crashes, requiring data such as driver demographics, road infrastructure, and weather. Viable solutions is the main work that we hope to do in future work, and future research in this database has endless opportunities to make a significant contribution to understanding and preventing car crashes in New York City.