INST 327 – Database Design and Modeling (Section: 0203)

5/15/23

Team 4: Philip Samuel, Malik Oumarou, Maria Gomes Master, Yash Gupta, and Zhen Zhou

Final Report

Introduction

Collisions, often referred to as accidents, are when 2 or more vehicles collide with other vehicles and/or people. Collisions are often inevitable. However, many times, they can be avoided. Collisions often are a result of at least one of many factors, such as distracted driving, drunk driving, and inexperienced driving. As a group, we want to research the cause, factors, and circumstances surrounding these collisions in New York City. The database we are using displays collisions which were reported by the New York City Police Department. Collisions are only reported if either there is an injury/death and/or damage needing at least $1,000 to repair are involved.

The dataset also consists of unique information related to each crash reported (within the dataset), such as collision id, time/date of collision, type of vehicle involved, make/model/year of vehicle involved, driver's license status, vehicle id, vehicle state registration, direction of vehicle traveling, number of occupants in vehicle, drivers' sex, pre crash information, point of impact, damage information, and public property damages information. The dataset is constantly being updated as more and more collisions occur in New York City. Hence, we will only be analyzing a certain number of accidents that occurred during a certain time period. We aim to use our dataset to bring awareness within society regarding collisions and how they can be prevented. Our college campus is an area where collisions could be prevented. We hope to advocate for safe driving practices using this dataset.

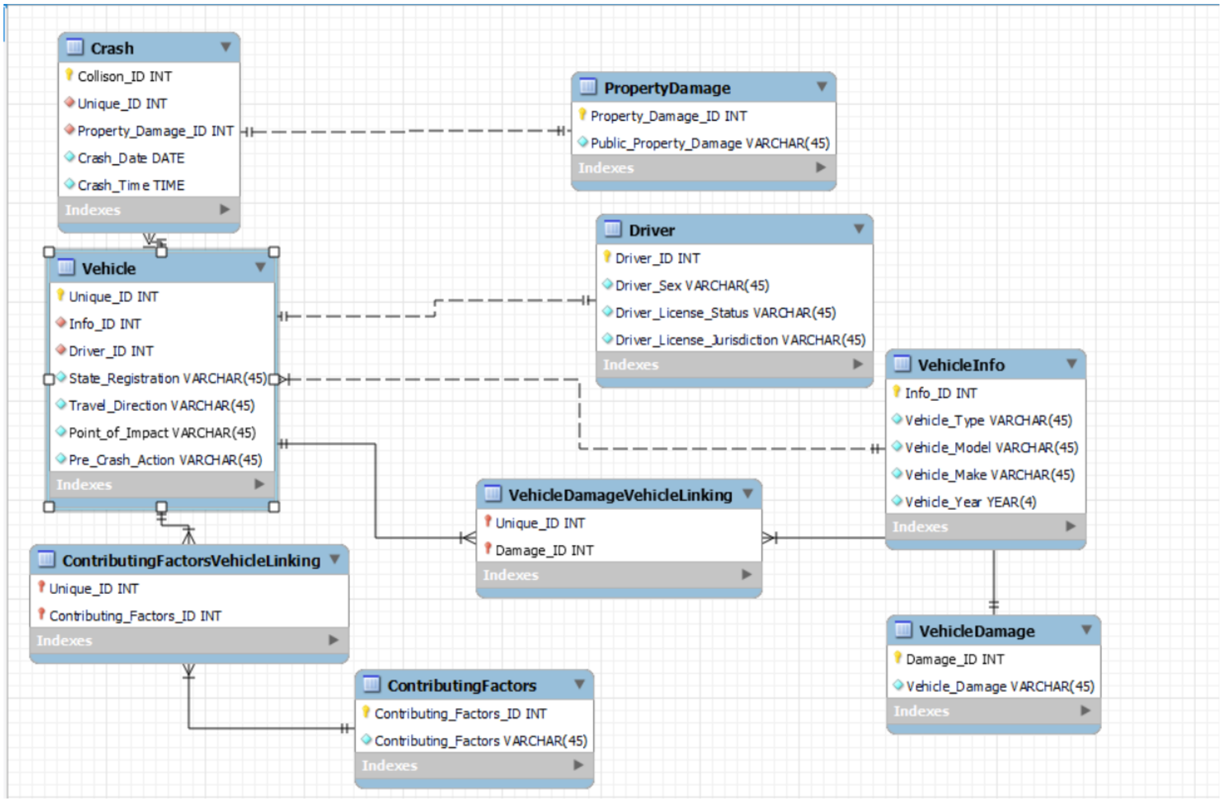
Database Description

The database focuses on the collisions within New York City. This database is reported by the New York Police Department, and includes information about collisions which occurred during a certain time period. Collision ID, date/time of collision, vehicle information, and much more are included in this database, which can be used to make conclusions or projections about collisions. We took out several rows of blank data, which we thought to be unnecessary to include. This could have been causing confusion.

The intended audience would be people who live in New York City. As seen in the database, the data only depicts collisions within New York City. Motorists can use this database to see potential dangerous intersections/areas within the city, where increased caution should be exercised while driving. Motor vehicle advocates/safety advocates can also use this database to see where collisions usually occur, and advocate for increased safety measures to take place in certain areas. They can also push lawmakers to increase law enforcement to be present in areas throughout the city, or to invest in safety-related technologies which can ensure and increase public pedestrian safety, as well as the safety of motorists. Lastly, pedestrians could use this dataset to advocate to lawmakers to make stricter laws benefiting pedestrian safety.

Figure 1

ERD Diagram



Sample Data Plan

Our team began to implement our strategically designed solution to find the leading causes of motor vehicle collisions in the state of New York. We used the State of New York's motor vehicle collisions dataset, available at this URL: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi nx95. From this data, so far we have filtered all the data we deemed unnecessary to our project from the dataset. We included data points most pertinent to the collision cases such as the vehicle's make, model, year, as well as when the date of the collision took place, and all other similar relevant data to the crash, the level of severity, and the drivers involved. Personal information was left out of the dataset. In addition, we have filtered down our dataset from around 1000 to about 100 rows, sorting by most recent collisions to maintain data relevance. The following are some samples of our first few tables in our logical design. The first table, CRASH, holds the values relating to the initial crash statistics, such as the time and date of the crash. The second table, DRIVER, holds the values relating to the statistics of the driver involved, such as their sex and driver's license status.

SAMPLE TABLE 1: CRASH

Collision_ID	**Crash_Date**	**Crash_Time**
4466407	10/12/2021	9:00:00
4466616	10/12/2021	16:30:00

SAMPLE TABLE 2: DRIVER

Driver_ID	**Driver_Sex**	**Driver_License_Status**	**Driver_License_Jurisdiction**
1	F	Licensed	NY
2	F	Licensed	NY

Changes from the Initial Proposal

Scope of the project

The scope of the project will continue to be the state of New York, because our database only contains data about motor vehicle collisions in New York.

Inclusion

After working through our project's logical design, through proper normalization and rigorous data analysis there is data which we decided to keep and also data we deem unnecessary. These new entities and attributes are following the 3NF.

“Table 1: Crash”. This table is just the “Accident Table” from our proposal which includes data regarding the date and time of the crash. The main change here is that we removed some attributes due to redundancy and its uselessness to our project's scope. The attributes of this table are: Collision_ID (PK), Crash_Date, and Crash_Time.

“Table 2: Vehicle”. This table contains specific information about the vehicle. The main change here is that we decided to not use the dataset’s Vehicle_ID, because of uncertainty of its values. We realized that Unique_ID can represent each vehicle since it is unique. Added Info_ID as it links to Vehicle Info Table, which contains information about the car specifications. The attributes of this table are: Unique_ID (PK/CPK), State_Registration (Varchar datatype), and Info_ID (FK), Point_Of_Impact (Varchar datatype), Pre_Crash_Action (Varchar datatype), and Travel_Direction (Varchar datatype).

“Table 3: Vehicle Damage”. We simplified this table to only have the primary key and the actual damage. Point_Of_Impact is moved to Vehicle Crash Linking Table because of its relationship between the car and crash. The attributes of this table are: Damage_ID (PK) and Vehicle_Damage (Varchar datatype).

“Table 4: Driver”. We decided to not include age as part of the privacy and lack of data. The attributes of this table are: Driver_ID (PK), Unique_ID (FK), Driver_Sex (Varchar datatype), Driver_License_Status (Varchar datatype), and Driver_License_Jurisdiction (Varchar datatype).

“Table 5: Property Damage”. This table contains data telling us if there is property damage or not. The attributes of this table are: Property_Damage_ID (PK), Unique_ID (FK), and Public_Property_Damage (Boolean datatype).

“Table 6: Contributing Factors”. This table specifies what contributed to the crash.

The attributes of this table are: Contributing_Factor_ID (PK) and Contributing_Factors (Varchar datatype).

"Table 7: Vehicle Damage Linking". Shows the relationship between Vehicle Table and Damage Table. The attributes of this table are: Damage_ID (FK/CPK) and Vehicle_ID (FK/CPK).

"Table 8: Contributing Factors Vehicle Linking". Shows the relationship between Vehicle Table and Contributing Factors Table. The attributes of this table are: Contributing_Factors_ID (FK/CPK) and Unique_ID (FK/CPK).

"Table 9: Vehicle Info". Has data dealing with the specifications of the car involved in the crash. The attributes of this table are: Info_ID (PK), Vehicle_Type (Varchar datatype), Vehicle_Model (Varchar datatype), Vehicle_Make (Varchar datatype), and Vehicle_Year (Varchar datatype).

Exclusion

The dataset of collisions in New York City underwent data cleaning before we normalized it. As there were over 100,000 rows of data present in the dataset, we have decided to work with a more reasonable amount of data for this project, around 100 lines. We will also be only taking into consideration data from November 2020 as it has the most amount of data for a full year. Additionally, we will only include rows of data in the dataset that match the attributes presented in our linking tables in the updated ERD we created. With these restrictions, we were able to exceed our goal of 100 lines (now 178 lines) of data in our cleaned dataset.

Many of the columns of data were empty as well. For instance, columns which represent vehicle data, driver data, and damage data. We want rows that do not have NULL values for most of the columns. Therefore, we have included rows that do not match the column criteria we mentioned above.

Given the focus on potential preventive measures analysis, we want to be able to exclude information such as:

Vehicle_ID. Initially, Vehicle_ID represented the number of vehicles involved in a crash (i.e. 1, 2, etc). Therefore, there were repeating values of Vehicle-ID for different collisions. Furthermore, it changed to larger alphanumeric strings (eg: b5527aee-ce69-497e-a842-4860d3cd14bf) which could possibly be the encrypted data for each vehicle for privacy issues. However, this alphanumeric string does not represent a vehicle's VIN. On the contrary, the Unique ID identifies each unique vehicle involved in a crash. Therefore, technically the Unique ID represents a vehicle. As a result, we will be *excluding* Vehicle_ID, and instead, use *Unique_ID*.

public_property_damage_type. The majority of the public_property_damage_type column for each collision was filled with null values. Furthermore, the ones that had values were descriptions of the type of public-property damage. Therefore, due to the inconsistency of data in the column, we will exclude it.

The dataset contains no confidential information about the driver themselves. For statistical analysis of a driver's demographic characteristics involved in a crash, a driver's sex (M or F), state registration (state abbreviation), and license status will be included. None of these attributes risk their privacy.

Overall, inconsistent, large amounts of missing data and variations in how the data was recorded will be excluded as it might be difficult to draw statistical inferences from such data.

Questions and information needs

Note: For the following 10 questions we wrote, our queries answered the first 5 questions. Each team member wrote one query and their names are listed next to their specific question.

1. Maria Master - In November 2020, which VEHICLE_TYPE committed the most crashes?
2. Yash Gupta - What is the safest Sedan to drive in New York, so that Point_Of_Impact is "No Damage"?
3. Zhen Zhou - Do drivers with a DRIVER_LICENSE_JURISDICTION outside of New York commit more crashes than those with one inside of New York?
4. Malik Oumarou - What is the leading contributing factor of collisions in New York in

November 2020?

5. Philip Samuel - For a Station Wagon/Sport Utility vehicle, what was the least occurring Pre-Crash Action?
6. Since the start of the Motor Vehicle Collisions - Crashes dataset (July 1st, 2012), which year had the highest number of crashes
7. What was the highest contributing factor for a motor vehicle collision in both the CONTRIBUTING_FACTOR_1 and CONTRIBUTING_FACTOR_2 columns of the dataset?
8. Were more crashes committed by licensed males or licensed females?
9. Does the number of VEHICLE_OCCUPANTS impact CONTRIBUTING_FACTOR_1 and CONTRIBUTING_FACTOR_2 of the crash or does it not matter?
10. Does POINT_OF_IMPACT have a significant impact on VEHICLE_DAMAGE?

Collaborations with the Professor, TA, and AMP

We would like to thank Professor Pamela Duffy, our discussion TA Nidhi Nambiar, and our team AMP Derek Miller for their tireless efforts in supporting our team as we complete this project. From the start of this project up until now, our team has been in constant contact with the instructional team and sought their advice many times. First, our team met with Professor Duffy to normalize our dataset for the first time via a Zoom meeting. Then, our team met with Nidhi and Derek several times to discuss normalization, datasets, and ERDs via Zoom meetings. We appreciate all their help and will continue seeking their advice for the remainder of this project.

Diversity, Equity, and Inclusion Considerations

After working with this dataset, which is designed by the New York Police Department, one can easily deem that diversity, equity, and inclusion factors were not taken into consideration when initially drafting this database. In this dataset, information such as vehicle information, gender of the driver, driver location information, etc. are given for users of the database to comprehend. However, this information has the potential to be misused, or be inaccurately depicted. This can cause errors when it comes to using this data for data reporting purposes. To elaborate, the only genders reported are "male" and "female". Although there is the possibility that everyone represented within the dataset identifies as either male or female, it is near impossible to say with certainty that there was nobody who identified with a gender other than male or female. Hence, the New York Police Department either deliberately discriminated against those who do not identify themselves as male or female, or there was actually nobody within the database who identified themselves as such. The reason why it is possible that this was done deliberately by the New York Police Department was because of the fact that the only 2 genders listed in the 2020 Census were "male" and "female". Hence, it is a possibility that the New York Police Department only considers 2 valid genders, which goes hand-in-hand with the 2020 Census.

In addition to this, the New York Police Department also only includes collision information from those which have occurred within New York City. In other words, collisions from other places neighboring the city have not been included in the database. This could cause issues when using this dataset for reporting, such as causing the data results to be one-sided or biased. People can make biased or judgemental conclusions about many factors, such as about the drivers in New York. Lastly, the dataset fails to include one important consideration, which is weather at the time of the collision. Weather conditions are very important to know when looking at collisions. This database fails to identify this information, which could be crucial for others who would want to deduce conclusions about collisions within New York City at the given period of time.

Data Privacy, Fair Use, Other Ethical Considerations

When designing a database with large amounts of public data, it is also important to consider the purpose of the database, its implications, and its potential impact on society. Furthermore, as there could be potential cases of data breaches and other mishaps, the access policies and security instructions imposed should be strictly maintained. Additionally, we must be aware of the potential risks of using data to make assumptions about individuals and communities, as this can perpetuate stereotypes and reinforce existing biases. This can also be influenced by the way the data was collected and recorded as well. Even though the data is being collected from an open source, it is significantly important to abide by the sets of rules and regulations placed under copyright or fair use laws. When working with large amounts of data, it is common to encounter missing or inaccurate data, which can affect the reliability and validity of our analyses. Therefore, we must ensure the quality of the data by cleaning it properly as well. Conclusively, while taking into account the ethical, legal, and social implications of our work, we should respect and consider the perspectives of diverse stakeholders as well. The protection of sensitive, confidential data and mitigating potential harm against our stakeholders will be at the forefront of our database design.

Lessons learned

After working with this dataset, we realized that data ethics are often overlooked when developing databases which are to be used by public audiences. We also realized that this could cause potential discrepancies when using that data for reporting purposes. Working around this could be cumbersome. Hence, we also learned how to work around ethical issues, especially by learning skills such as cleaning up databases, normalizing data, and much more.

We also learned that it is often possible to have multiple answers to questions, relating to the database. Hence, we as a team realized that it is important to use solutions that efficiently and effectively answer the questions. For example, some of us developed queries to the same question. We put ourselves in the shoes of those who may actually use this database and see which query is appropriate for answering the question.

Potential future work

Many improvements have been made to the dataset ever since we worked on it from the project duration. Our team members worked with the instructional team of INST 327, and with their assistance, we were able to efficiently clean up our dataset and use it to make queries which could be (and were) used to answer numerous real-world applicable questions. We initially had a goal of having 100 rows of data, and we actually ended up with over 178 rows of data! Using the strategies used during this group project, we hope to utilize those techniques and apply them to similar databases, and be able to use them for a public good. If we decide to make improvements and continue to use this database, we could add columns which would act as clarifying factors, such as color of car, weather at the date/time of collision, and many other factors. These factors are not currently present within the dataset, and could help those like us who use the dataset to deduce conclusions about collisions.

Views

View name **(team member name)**	**Req. A - JOIN**	**Req. B - Filter**	**Req. C - Aggregate functions**	**Req. D - JOIN linking + base table**	**Reg. E - subquery**
stationwagon_pre_crash_action (Philip Samuel)	X				
collision_contributing_factors (Malik Oumarou)	X			X	
driver_crash_counts (Zhen Zhou)	X	X	X		
safe_vehicles (Yash Gupta)	X	X			
crash_damage_overview (Maria Gomes Master)	X	X	X		X