NST327: Database Design & Modeling
Department of Information Studies, University of Maryland, College Park
Team Project Final Submission (Database + Report) – May 15, 2023
Section 0201 – Group 12
Zachariah Zachariah

Team Project Final Submission (Database + Report)

Introduction

Our database will be based on the Motor Vehicle Collisions dataset. For this database, our team decided to pursue the domain of what factors such as vehicle makes, states, regions, times, and days, months, and years contribute the most to vehicular collisions. Here, we will be looking at the attributes and rows from the year 2016. The main attributes that will be considered for this database include attributes pertaining to vehicle collisions, driver, vehicle, damages, and contributing factors. In addition, we will only focus on three northeast states, which are Virginia, Maryland, and the District of Columbia. These states collectively are locally known as the DMV.

One of the groups of people that the database is intended to benefit is new vehicle buyers living in the DMV who can use the database to determine which vehicle makes have the most collisions in a given year, so they can make the best choice when buying a vehicle. For example, new vehicle buyers may want to avoid vehicle makes that are involved in the greatest number of crashes which they can find out using our database. Another group would be people who would want information that can help them keep their probability of being in an accident low when taking into account factors such as the time of crash and vehicle make in which the crashes took place. Lastly, another group that would benefit from using our database is insurance companies, as they can use attributes like what state had the most crashes and what vehicle makes are involved in the greatest number of accidents in a given year to calculate and charge insurance premiums for new customers. For example, if the database showed that Maryland state had the greatest number of accidents, they could charge higher premiums to new customers residing in the state of Maryland, helping the company make correct financial decisions.

Database Description

Logical Design

For the logical design, we first created fields from the Motor Vehicle Collisions dataset that we thought would be important for successfully serving the purpose of our database. The

fields that we thought would be the most important were the ones that contained specific information about the vehicular crash, such as crash time and date and as well as vehicle information. After choosing the fields, we created tables where they could be added too. The tables were named in a way where a user can easily understand what attributes and information the table will contain for quick access and retrieval. For example, the Collisions table was created so that any information pertaining to particular vehicular collisions can be accessed. Thereafter, we started the normalization steps for the tables, where we were able to remove repeating groups and partial and transitive dependencies. Lastly, we connected all the tables by their primary, foreign, and composite keys, which showed the correct relationships the tables had, which consisted of one-to-many and many-to-many relationships.

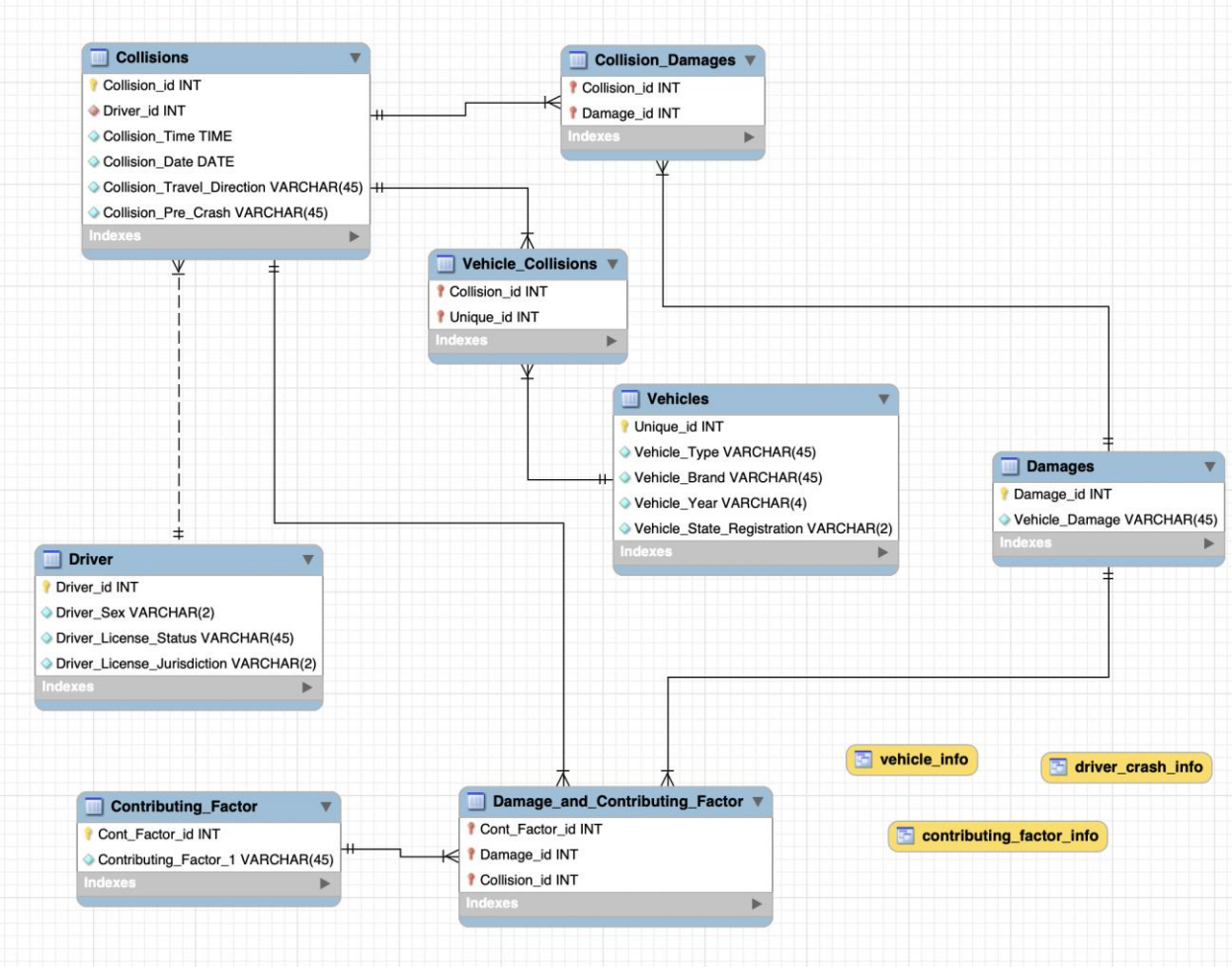

Image 1. Entity Relationship Diagram (ERD)

Physical Database

For our final physical database, we had a total of eight tables. We had three tables that were our join(linking) tables which were Vehicle_Collisions, Damage_and_Contributing_Factor, and Collision_Damages. The linking tables all had about 20 rows of data. On the other hand, our non-join tables were Driver, Collisions, Contributing_Factor, Damages, and Vehicles, where they had between 14 to 20 rows of data. All these tables in the database would help answer questions related to vehicular collisions that occurred in the year 2016 and took place in vehicles registered in the states of the District of Columbia, Maryland, and Virginia. For example, from our database, we wanted to give users information about the type, brand, and year of vehicles that were involved in crashes and the contributing factor for the crash, which is why we added Vehicles and the Contributing Factor tables. We added all the other tables with the same thinking process in mind.

Sample Data

After revising our database plan, our team realized that our selected database would be too large to handle based on the data we initially decided to have in our database. Therefore, we added new constraints to help make our database smaller and more concise. Our first constraint is that we should only have data on vehicular crashes that took place in the year 2016. However, we decided to add another constraint which is to limit the states we will include for the vehicle crashes. We decided to include only three states from the northeast, which are the District of Columbia, Maryland, and Virginia, as initially, we decided to include all 14 states in the northeast. Therefore, the sample data of the attributes that were affected due to these constraints in our database are 'Collision_Date', 'Vechicle_State_Registration', 'Driver_License_Jurisdiction', and 'Vehicle_State_Registration.' The other sample data will include data about the driver, such as driver sex, driver state, and the vehicle, such as make and model, the damage the vehicle incurred, and details about the collision, such as time and date.

Views / Queries

View Name	Req. A	Req. B	Req. C	Req. D	Req. E
Driver_Crash_Info	X	X			
Vehicle_Info	X	X	X	X	
Contributing_Factor_Info	X	X	X	X	X

Toward the end of the project, we made three queries which we saved as three separate views with names that describe what information the query would be displaying. The first view/query, Driver_Crash_Info, is intended to give driver information and collision information for vehicular collisions. More specifically, it gives information about the driver's sex and the state they are licensed in. The view/query can help a user look for information such as which driver sex is involved in the most significant number of vehicular collisions. There is also information about the time the collision took place and the date. The second view/query is called Vehicle_Info, which gives information about vehicle types made after 2010 and how many there were in total. Here, users can find information about what vehicle types past 2010 are involved in the greatest number of crashes. Lastly, the third view/query is called Contributing_Factor_Info, which presents information on what contributing factors are the most prominent in vehicular collisions.

Changes From Original Design

Our final database saw some major changes from our last change in design, which was made during the project progress report assignment. One of our original major idea that we wanted to implement in our final database was to have data from all northeast states. However, we had realized that the data was too enormous and complex for us to handle. Therefore we decided to shrink the states to the states in the DMV, which are the District of Columbia, Maryland, and Virginia, which was also motivated by the fact that we live in the DMV area. Another major change was to reduce our original range of the years from 2015 to 2019 to the year 2016. We decided to make this change because the data from the original range was too large for us to transform and clean as the data ranged in hundreds of thousands of rows. Therefore, we chose the year 2016 as it had way less data but enough data to meet the purposes of our database, one of which is having collision reports from recent years than from a decade ago, where the latter would not provide much helpful information for our target audience in the present day. However, everything else will remain the same as initially presented in the proposal and the project progress report, where we will be looking at factors such as vehicle makes, states, time, and year that contribute to the most vehicle collisions.

Database Ethics & Data Privacy Considerations

The presentation on Database Ethics had profoundly impacted our project. The presentation mainly affected the sample data we had included. From the presentation, we had

learned that data can often unknowingly have biases and copyright concerns, which can raise ethical concerns, and data that reveals too much information that raises privacy concerns. Therefore, our team decided to look for any possible ethics and privacy concerns that the sample data we plan on using would pose for our final database design. One of the potential privacy concerns we see with our database is displaying the driver's sex, license status, and license jurisdiction. We do have a concern about whether or not this information can be used to retrieve more personal sensitive information about the driver, such as their name or address. Another concern we have is pertaining to copyright concerns since we are using the brand and model names of vehicles of major car companies. These brand and model names are copyrighted, and we to make sure that we do not use the data in a way that results in a copyright violation. Other than these concerns, we do not see any other potential database ethics and data privacy concerns with our final database.

Lessons Learned

One of the biggest issues we had with the project was the lack of knowledge of certain topics presented in class that were required to finish project deliverables. An example of this is during the project's logical design submission. The entire group was confused about how to perform the normalization steps on the sample data we had collected. We were stuck on this project deliverable for a couple of days, and then we resolved this issue by deciding to arrange a meeting with our AMP and section TA so we could get further assistance. Another issue we had was a technical issue using MySQL workbench to import our sample data to the tables of our created database. We were able to import data into some of our tables but were not able to import data to our join (linking) tables, and after many attempts of searching online and through the class slides, we were unable to import the data. Therefore, we decided to ask for help during office hours. Here, we worked together with the TA and found that the data in our join (linking) tables had the wrong format, which SQL does not recognize. Thereafter, we changed the affected rows of data into the correct format and resolved our issue. The TA also taught us how to use the activity log in the import wizard to figure out what errors are occurring when trying to import tables which also helped us solve any issues that appeared during data import. Based on the issues we have mentioned, one of the lessons we learned is not to be afraid of asking for help from the instructional team when we are stuck on something, whether it be issues between team members, issues with the project, or technical issues.

Potential Future Work

In terms of potential future improvements and extensions to our database, we would add all the information that we decided to eliminate due to the short time we had since the data for all that information would take a hefty amount of time to import, clean, and transform. For example, we would increase the range of years of the data for the database and possibly include all the US states so our database could be inclusive and beneficial to everyone in the country. Another addition I wanted to add to the database is information about seasons, as they play a significant role in vehicle collisions, as the summer season is known to have the most significant number of crashes (*When and where most accidents occur*, 2021). I could have added this attribute since we had information on the crash dates, but it would have taken some time to identify and categorize the crash dates by season. I would also like to incorporate the database into an actual product, such as a website or app where API is used to collect information from our database for the product.

References

When and where most car accidents occur. Law Offices of Marion M. Moses, LLC. (2021, July 1). https://www.moseslawsc.com/blog/2021/july/when-and-where-most-car-accidents-occur/