INST327, Section 0101

Final Project Report

December 5, 2023

Team 6

Ahmed Elhag, Joseph Rohan Kettish, Zayd Mahfuz, Antoine Williams, Manuel Orallo

Introduction

The "Motor Vehicle Collisions" dataset, sourced meticulously from the New York Police Department, spans the years [2012-2021]. In navigating the vast terrain of driving behaviors, our focus centers on the distinctive period of [2018-2021], a time marked by the peak of the COVID-19 pandemic. This project aims to dissect and analyze the nuances within crash data during this pivotal timeframe, shedding light on the evolving dynamics of vehicular incidents. Comprising 25 columns, each detailing various aspects of vehicular collisions, this dataset is a comprehensive resource.

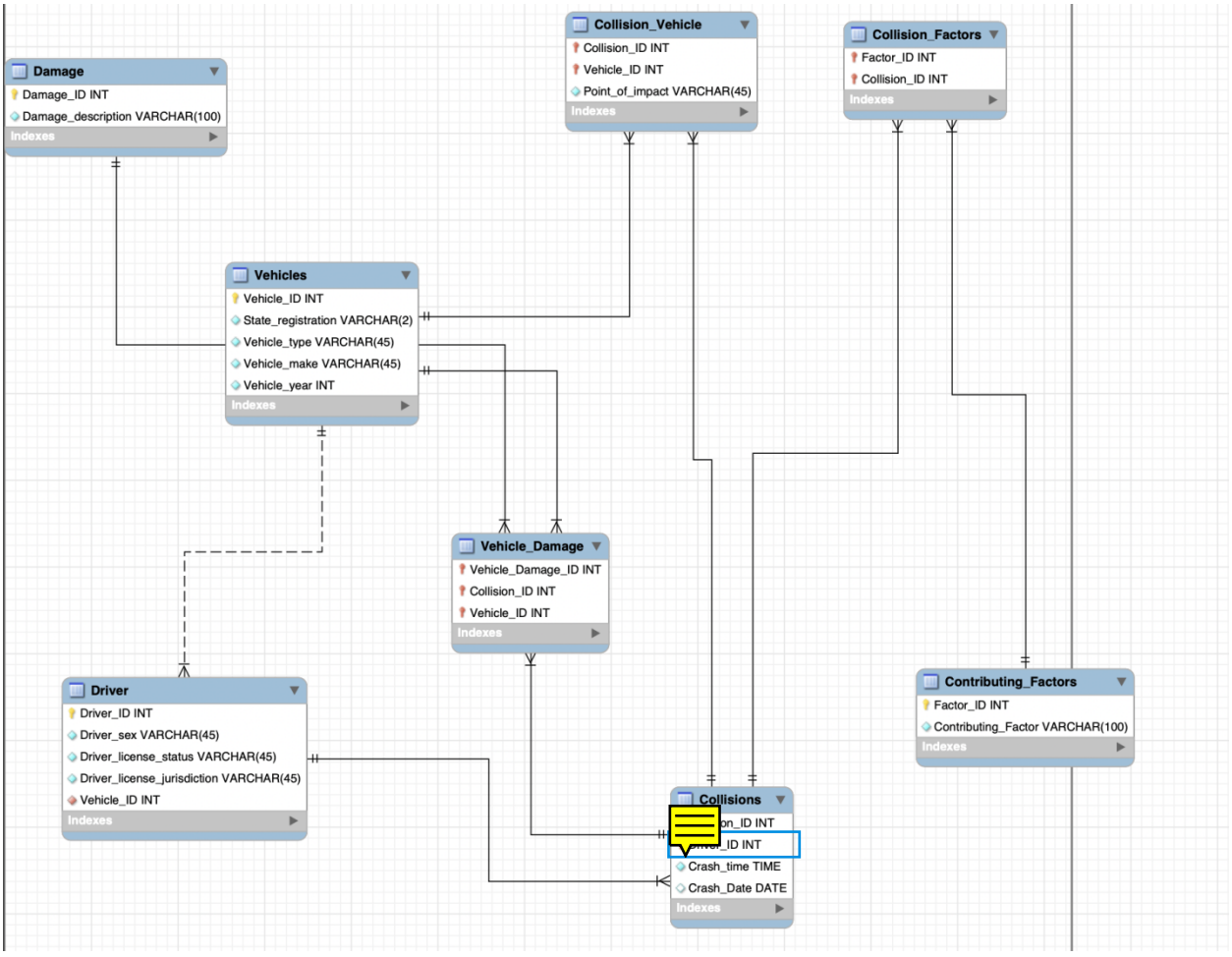
Our examination extends beyond mere statistics, delving into temporal considerations, crash severity, and the identification of repeat offenders. Noteworthy is the dataset's ongoing relevance, receiving frequent updates and remaining an active tool for the city. The rationale behind our chosen topic lies in the ubiquity of driving experiences, irrespective of one's level of expertise. Motivated by a shared interest in dashcam videos, our team collectively decided to delve into the repercussions of the pandemic on driving habits. The distinctive backdrop of New York City, renowned for its dense population, adds an extra layer of complexity to our investigation. This exploration is not just about numbers; it's an inquiry into the impact of unprecedented events on our shared urban experience.

Database Description

The “Motor Vehicle Collisions" data set is from the New York Police Department and ranges from 2012-2021. To pare down the set, we decided on a time frame, which led us to the idea of looking through the crashes during the pandemic. For this project, we have decided to focus on “How does the crash data vary from 2018 - 2021?”. We chose this period as it was the height of the COVID-19 pandemic. Most of the attributes we will store are related to the time of the crash and severity; we will also account for repeat offenders. The dataset has 25

columns, each regarding a detail of the vehicle collision, such as what the car was doing pre-crash. The data is also updated frequently and is still being used by the city.

The main reason we have chosen this topic is because driving is something we can all relate to, whether we're just starting or are experienced. Some of our group members watched dashcam videos, so they suggested the topic; after some thought, we all went with it and noticed that we could research the pandemic's impact on driving. The data being from New York City makes this more interesting, as the city is very densely populated. Furthermore, according to NYC, the reports the dataset holds are for collisions where someone is injured or killed or where there is at least $1000 worth of damage.



Sample Data Plan

When beginning the project, we decided to focus on the motor vehicle collisions reported by the New York Police Department (NYPD) during 2019-2021, which is also known as the height of the COVID-19 pandemic. Our overall question of "how did the pandemic affect driving" has stayed the same throughout our revisions. This allows us to include the required number of tables this project needs while also narrowing down the amount of data we have by taking away tables that weren't necessary. The information we plan on filtering out is any information outside of the time frame.

Logical Design/Examples

1. Collisions Table
    - COLLISION_ID (Primary Key)
    - CRASH_DATE
    - CRASH_TIME
    - Driver_ID (FK)

COLLISION_ID (PK), integer	CRASH_DATE (date)	CRASH_TIME (time)	Driver_ID int
100201	2019-07-09	9:03	1
297666	2020-04-25	21:15	2
3308693	2020-04-10	20:34	3

2. Table Vehicles

Vehicle_ID	State_Registration	Vehicle_Type	Vehicle_Make	Vehicle_Year

1	NY	Pick-up Truck	NISS - CAR/SUV	2011
2	NY	Sedan	VOLK - CAR/SUV	2012
3	NY	4 dr sedan	NISS - CAR/SUV	2015
4	NY	Sedan	KIA - CAR/SUV	2015

3. Vehicle_Damage

Vehicle_Damage_ID	Collision_ID	Vehicle_ID
1	3831473	1
1	3831473	2
3	3734552	3
6	3838743	4

Views/Queries

View name	Req A. (2+ tables) Need 4 Done	Req. B (filtering) Need 3 Done	Req. C (aggregate) Need 2 Done	Req. D (linkint table) Need 1 Done	Req. E (Subquery) Need 1 Done
fact_gt_Y2k	X	X		X	

Car_by_Time	X	X	X	X	X
max_time	X	X	X	X	
Maximum_dam	X		X		
Driver_vehicles	X				

Progress Report:

Since completing our proposal and logical design, we have received a lot of feedback regarding our normalization structure as well as some feedback regarding our ERD. To improve our normalization structure, we consulted with the TA to see where we could make changes, which was mostly in the 1NF and 2NF steps. We then updated our normalization structure, and this ultimately helped us see where we needed to fix our ERD. For our ERD, we updated the overall structure, including certain relationships and entities. We understood that this was a crucial part of making our dataset accurate, so we incorporated all the feedback given and updated it to our best understanding. As a team, we worked together to update our normalization table and our ERD.

Changes from the Initial Proposal

The scope of our project remained the same; we plan to focus on “How does the overall collision data vary from 2018 - 2021?”. We chose this period as it contains not only the height of the COVID-19 pandemic, but also the last year of pre-pandemic Earth. We believe that the pandemic has had an impact on every individual, so we were interested in how it affected drivers. Some questions we think our project could answer is seeing how many accidents occurred during the height of COVID-19 and whether a specific vehicle type was involved in the majority of collisions. This can help fulfill the

information needs of individuals who want to know more about vehicle collisions and the safety of a particular vehicle. It can also help the New York Police Department analyze the effect a pandemic has on a city as densely populated as New York City and take note of what causes these collisions (pedestrians, bicycles, other vehicles, etc.). The data can also be used to help improve safety for certain vehicle types and take further measures to avoid collisions altogether in the future. Overall, our project has stayed the same, with the only change being the removal of the Property Damage Table, as the original dataset had numerous blank (null) entries during our time frame.

Diversity, Equity, and Inclusion Considerations

We have changed a couple of things since we first made the proposal. There are some things that we like to keep the same, such as the point of bias being the gender of the driver. To keep it non biased, we will include any and all of the crashes, despite the gender. Another topic that could be biased as well is the time of crashes. Nighttime crashes might be overrepresented, so ensuring we include more crashes during the day will make it more unbiased. By making these types of changes and a few other things, we can make the dataset more diverse.

Ethical Considerations

Since the start, even when we worked on the proposal, data privacy has been a priority to us, as the dataset contains personal information like names and addresses. We have since decided to remove that information, as it's important to know the details of each crash without publicizing those involved. That information is useful to the police, but not to the public, as it might lead to ostracism in these drivers' communities. Based on the dataset that was given to us, not all the aspects and factors need to be included. If we were to base it on the original report, we would leave out fatalities and injuries, as that might be too sensitive to include. We will still like to safeguard the dataset to prevent unauthorized access and potential breaches.

Lessons Learned

We started off this project with a baseline understanding of SQL but no technical foundation for how to practically apply it. Over the last 14 weeks, we've learned a variety of concepts regarding SQL and its queries. This spans past the base foundation and encompasses more in depth material such as Aggregate Functions, Complex Queries, CTEs, etc. Our understanding of how to develop and execute SQL queries has allowed us to better grasp concepts like database management and structure. Another thing that we learned throughout the project process was about ERDs. This was something in our project that we had to get fixed early on when working on it. We consulted with a TA in order to receive feedback on it, and we ended up fixing our mistakes and having our ERD where it is supposed to be. The feedback that we had was to connect any tables that did not have relationships by adding foreign keys.

In addition to technical skills, we learned integral concepts such as normalization. At first, this concept was hard for the team to wrap our heads around. However, through labs, office hours, and practical exercises, we began to develop an understanding of this concept and its necessity. Normalization allows one to take an unorganized dataset and eliminate redundant data, minimize data modification errors, and simplify the querying process. We were able to apply this concept in the creation of our ERD model. Normalization also provided us with the ability to find and produce linking tables to develop a coherent and connected database. Creating views was another concept that we were on the fence about; however, the lab session where we did some exercises on it helped a lot and impacted our success on the project. The views helped a lot in making viewing the database simpler. We also learned how to create database backups during lab sessions, and it simplified how to do it.

Potential Future Work

There are plenty of ways that this dataset can be used for potential future work. An idea for future potential work can be investigating the temporal patterns of collisions during [2019–2021]. Examining whether there are particular weeks, days of the week, or hours of the day when the accident rate is higher? Understanding

temporal patterns can help determine when accidents are most likely to happen. We can also investigate the differences in accident frequency and severity between the pre-pandemic, pandemic peak, and post-pandemic periods. This can help demonstrate how driving habits and crash patterns varied throughout the pandemic. The severity of accidents can also be used to investigate whether certain types of vehicles are more prone to causing or being involved in severe accidents. Analyzing the injury and fatality rates can provide insight into the safety of different vehicle types and where improvements can be made. Using the dataset's location can also allow us to apply geospatial analysis. This can help determine which intersections or locations have a high collision frequency. Enhancing traffic safety and urban planning can both immensely benefit from this information.