Bereket Asrat, Kyrene Jamero, Samnorp Deung, Wesley Cai
INST 327 - 0101
Team 6
09 Dec 2024

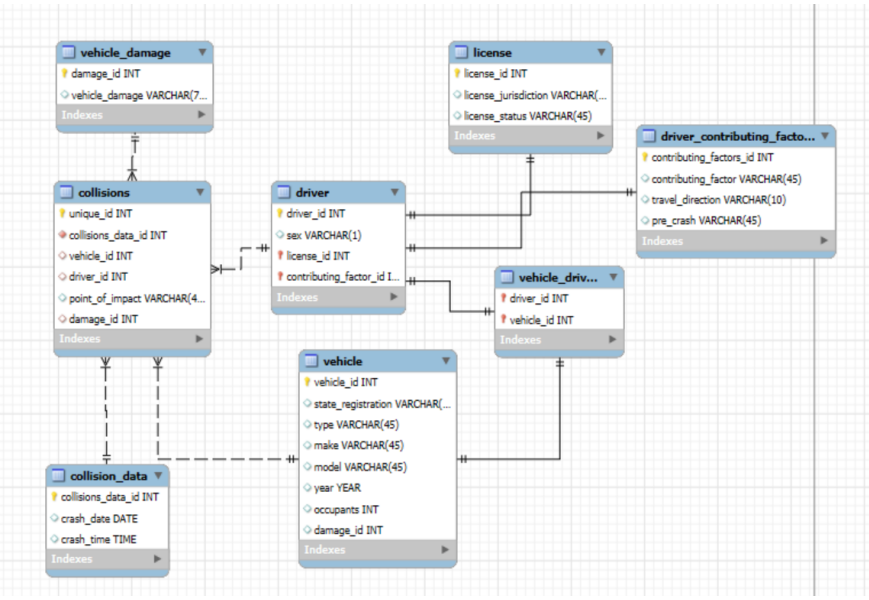
Project Report

Introduction/Database Description:

Alcohol related motor vehicle incidents pose a major threat to public safety, contributing to injuries, fatalities, and property damage every year. Thousands of incidents are reported every year, highlighting the urgent need for actionable insights to reduce and mitigate their frequency and impact. Our project addresses this issue by creating a relational database that integrates and analyzes alcohol related motor vehicle incidents, utilizing the NYPD's publicly available dataset from 2012 to 2021. This dataset provides detailed records of alcohol related motor vehicle incidents resulting in injuries, fatalities, or property damage exceeding $1,000 threshold. By structuring and normalizing this data, the database enables an in-depth analysis of key factors, like high-risk times and locations, driver demographics, and crash severity contributors. Key features of the database include the ability to explore relationships between variables, like driver demographics and crash severity, and to analyze trends over time. Data normalization ensures data consistency, reduces redundancy, and it improves accessibility for users.

The main goal of this database is to provide data driven insights to policymakers, law enforcement, and public health officials, empowering them to implement targeted interventions. For example, stakeholders can identify high risk behaviors and peak accident times, enabling effective resource usage and targeted awareness campaigns. The intended audience is policymakers, law enforcement agencies, and public health groups. By providing reliable data and meaningful insights, this project equips stakeholders with the tools needed to develop effective strategies, fostering safer roads and reducing the harmful impact of alcohol related motor vehicle incidents in New York City.

Logical Design:



Our logical design aimed to be intuitive and functional for users who want to learn more about the statistics of alcohol related motor vehicle accidents. After long hours of brainstorming, we came up with eight tables to structure our database with so that our information can be organized. This arrangement also allows us to write queries that can answer the questions we designed our database to answer.

The relationships between the tables are mainly one to many or one to one. An example of a one to many relationship is that one driver can be in multiple collisions, while an example of a one to one relationship is between a license and a driver because only one license can belong to one driver. Our group thought it was necessary to separate driver_id and vehicle_id into their own table because a driver can only drive one vehicle in any collision, but that can happen for multiple collisions. This way, our queries can be more accurate and no information will be lost or duplicated.

Physical Database:

Our database has the goal of having answers to questions about alcohol related motor vehicle accidents. This data was gathered from the New York Police Department, which included data about every collision that has been reported to the police over the past decade. Due to the massive amount of data, we decided to only include data that was in the most recent three years. We also omitted data about collisions that included drivers that were not affected by alcohol because that does not align with the purpose of our database. The collisions and

collision_data were separated into different tables because one collision affects two or more drivers in different ways. To accommodate this, collision_data represents data that can be applied to the collision generally such as the date and time of the collision. The collisions table represents what drivers and vehicles were involved in a specific collision, which is why they needed to be separated into different tables to avoid confusion. Tables that represent different entities so that queries can address specific questions about them were made such as driver, license, vehicle, and driver_contributing_factors.

Sample Data:

Collisions Table

Unique_id	Collision_ID	crash_date	crash_time	State_ registered
10385780	100201	9-7-2012	08:13	NY
17044639	3433415	5-2-2016	17:35	NY
14809587	3268858	8-1-2015	08:17	NY

Vehicle Table

Unique_ID	Vehicle_ID	vehicle_type	Vehicle_make	Vehicle_year	Travel_direction
17044639	219456	4 dr sedan	MERZ -CAR/SUV	2015	East
17211986	513371	Van	FORD	2010	East
16952374	87497	Station Wagon/Sport Utility Vehicle	TOYT -CAR/SUV	2015	South

Driver Table

Unique_id	Vehicle_ID	Driver_ liscence_status	Jurisdiction	Driver sex
17303317	672828	Licensed	NY	F

19139721	5bb0b59a-ce74-4a04-9f92-1446ebfe4f46	Licensed	NC	M
19139740	775bcb70-da96-4f01-bf71-f4b0cc3a4428	Permit	NY	M

Views and Queries:

Query Name	Join	Filter	Aggregate	Linking	Sub-Query
most_common_incident_times	x	x	x		
most_common_vehicle_types	x	x	x		
driver_sex_incident_count	x	x	x		
top_contributing_factors	x	x		x	
average_vehicle_age_incidents	x	x	x		x
Total	5	5	4	1	1

The following list describes what each query we wrote for our database displays:

Query 1: Displays the most common times and days for alcohol-related motor vehicle incidents.

Query 2: Displays the most frequent vehicle types involved in alcohol-related incidents.

Query 3: Displays the number of alcohol-related incidents categorized by driver sex(male or female).

Query 4: Displays the top contributing factors for alcohol-related incidents, grouped by travel direction.

Query 5: Displays the average age of vehicles involved in alcohol related incidents, with a subquery filtering vehicle ages.

Changes from Original Design:

An alteration which we have deemed necessary includes adding more details about the vehicle. Initially, we chose to focus on a more limited scope in which more significance was placed on the context of the accident–like where the vehicle was traveling–instead of the specifics of the vehicle involved. This was done because we initially concluded that details about the vehicle were unnecessary given our goal outlined in the project proposal, which was to create a rough profile for a drunk driver in order to aid understanding in drunk driving for improved regulations. However, it was decided that this data was ultimately relevant because not only could the difference in choice of vehicles establish an effective profile, but including this data could also eliminate a possible bias. Because the database only involves cases which meet a certain threshold for property damage or individuals injured, it is important to include the specifics about the vehicle since the level of damage to both property and individuals depends in large part on the make and model of the vehicle involved, as structures differ vastly between different vehicles. If this key factor in the severity of the incident, as well as if the incident made it onto the data at all, was omitted then a key part of the context would be lost; leading to potentially inaccurate results.

Ethical Considerations:

Our comprehensive database correlates driver demographics and vehicle specifics to analyze complex patterns across different communities. By examining multiple variables, we'll identify factors contributing to critical risk assessments and underlying patterns. The research will include balanced data collection to minimize sampling biases and ensure representative analysis. Driver privacy remains paramount, with personal details anonymized through random IDs and strong security measures like encryption and restricted database access. Data collection will use broad categorizations such as neighborhood-level locations and hour-rounded timestamps to prevent individual identification. The findings will support targeted interventions, focusing on constructive solutions and policy improvements rather than unfairly targeting specific groups. Ultimately, the database aims to promote comprehensive understanding by developing evidence-based strategies.

Lessons Learned:

Our team learned many lessons while developing this database. Early on, we adjusted our project scope based on graders feedback, recognizing the importance of including vehicle specific details like make and model. This change allowed for a more comprehensive analysis of accident severity and highlighted the need for flexibility when addressing project objectives. Data normalization presented a lot of challenges. Addressing inconsistencies, such as missing values in driver demographics and ensuring accurate data formats took a lot of time, practice, and effort. This process highlighted the importance of attention to detail in creating a working database. Effective communication and task management were important throughout the project. Weekly check-ins and task tracking ensured all assignments were done on time, despite managing individual responsibilities. These steps helped us stay organized and ensure timely completion of tasks while maintaining a smooth workflow. Finally, the Database Ethics presentation played a significant role in shaping our approach. It underscored the importance of presenting findings in a way that avoids reinforcing biases or stigmatizing specific groups. This helped guide how we structured our database and framed our analyses, focusing on valuable insights, promoting fairness, and ensuring it remains a reliable resource for stakeholders to use.

Potential Future Work:

Future improvements to the database could include expanding the timeline beyond 2021 since we are in 2024 or adding data from other cities to allow for a larger trend analysis. Integrating weather conditions, like rain or snow, would provide useful insight into environmental factors and how they affect crash severity. Custom views which target specific trends like vehicle types or driver demographics would enhance the database's usability further. One more thing we could do is develop a database app with features like push notifications which could improve accessibility and support real time decision making. Since the database is intended to influence policy making, diversity and inclusivity ethics dictates that community leaders, both private citizens and civil servants, must be consulted for any policy influencing use of the database. We have identified this as a requirement as to not unfairly implicate any community or group unfairly.

Conclusion:

This project successfully created a relational database to analyze alcohol-related motor vehicle incidents in New York City from 2012 to 2021, providing valuable insights into trends like high-risk times, locations, and contributing factors. By normalizing the data and examining ethical considerations, the database ensures reliability, and fairness. The database provides stakeholders like policymakers, law enforcement, and public health officials with insight into how to design targeted interventions and improve road safety. This project provides a strong foundation for ongoing efforts to reduce the harmful impacts of alcohol-related motor vehicle incidents and create safer communities.

References

New York Police Department. (2024, November 20). Motor vehicle collisions - crashes: NYC open data. *NYC Open Data*. Retrieved from https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Centers for Disease Control and Prevention. (2023, June 1). Impaired driving: Get the facts. Retrieved from https://www.cdc.gov/motorvehiclesafety/impaired_driving/index.html

National Highway Traffic Safety Administration. (2024, August 1). *Traffic safety facts 2022 data: Alcohol-impaired driving.* National Center for Statistics and Analysis. Retrieved from https://rosap.ntl.bts.gov/view/dot/78162

Mothers Against Drunk Driving. (2023). Drunk driving statistics. Retrieved from https://madd.org/statistics

National Weather Service. (2024). Historical weather data for NYC. Retrieved from https://www.weather.gov