INST 327 (0103)

FINAL REPORT - 05/11/2024

GROUP: 0103-G1

GROUP MEMBERS: Gabriel Gonzalez, Weizhen Ye, Johnathan Hill, Mohammad Ibrahim, Motunrayo Aduloju

Introduction

This project analyzes the Motor Vehicle Collisions dataset from NYC Open Data to gain insights into police-reported collisions in New York City. By selecting relevant columns and leveraging the dataset's 2 million rows and 29 columns, we aim to provide actionable recommendations to improve road safety and reduce traffic-related injuries. Our analysis considers urban infrastructure, traffic patterns, and historical data to comprehensively understand the factors influencing collisions in NYC.

During the logical design stage, we selected essential columns like Unique ID, Collision ID, Crash Date, etc., to capture each collision event's key aspects. We used linking tables to establish clear relationships between entities and minimize data duplication. The Entity-Relationship Diagram (ERD) helped visualize the database structure and entity interactions. In the physical design stage, we implemented the logical design in MySQL Workbench, creating tables based on the ERD entities and defining appropriate data types for each attribute. We established primary and foreign key constraints to ensure data integrity and maintain table relationships. The physical design process involved iterative refinements based on instructors' and TAs feedback and our observations.

Database Description

Our analysis is centered around the Motor Vehicle Collisions dataset sourced from NYC Open Data, focusing specifically on incidents within New York City. By selecting relevant columns like Unique ID, Collision ID, Crash Date, and others, we aim to explore the detailed aspects of police-reported motor vehicle collisions across the city. With a vast dataset, our analysis holds potential to uncover insights across various dimensions of collision events.

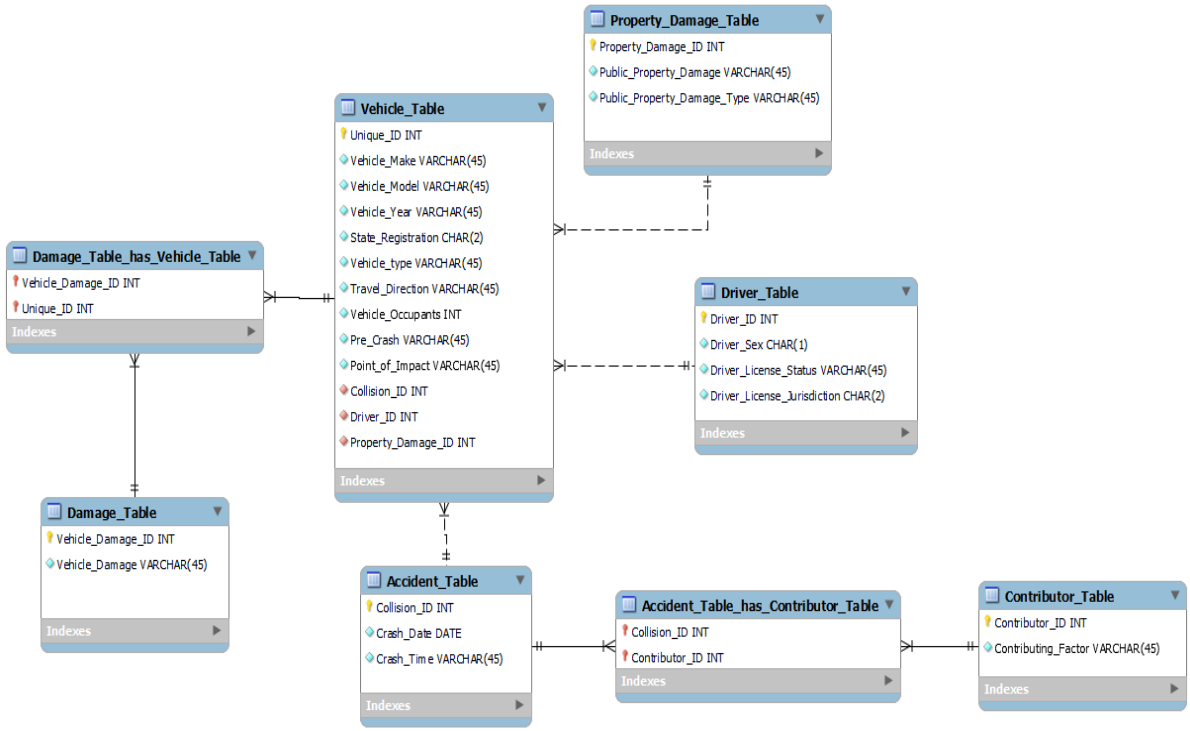
Logic Design

INST 327 (0103)

FINAL REPORT - 05/11/2024

GROUP: 0103-G1

GROUP MEMBERS: Gabriel Gonzalez, Weizhen Ye, Johnathan Hill, Mohammad Ibrahim, Motunrayo Aduloju



Our primary goal is to utilize this extensive data to develop actionable recommendations aimed at enhancing road safety and reducing traffic-related injuries. To achieve this objective, our analysis encompasses multiple dimensions, including urban infrastructure, traffic patterns, and historical data. By examining longitudinal trends and evaluating the effectiveness of past and present safety measures, we seek to provide a comprehensive understanding of the factors influencing traffic collisions in NYC.

We also incorporate additional linking tables. These tables contribute to a deeper understanding of data relationships and help mitigate the risk of data duplication. A crucial aspect of this integration involves combining our Vehicle table with the Collision table, optimizing our data structure for more effective analysis.

Physical Database

INST 327 (0103)
FINAL REPORT - 05/11/2024
GROUP: 0103-G1
GROUP MEMBERS: Gabriel Gonzalez, Weizhen Ye, Johnathan Hill, Mohammad Ibrahim, Motunrayo Aduloju

The database we're using is designed to meet the needs of UMD affiliates by providing detailed information about vehicle accidents. We started by shaping the physical design of our database based on the data retrieved from ELMS. We included the results that the dataset provided and considered the needs of our users when incorporating the data. For example, users might want to know which types of cars are involved in the most accidents or the most common reasons for accidents. To address these needs, we created eight tables to organize the data effectively. After importing the data into our ERD design and database, we added a total of 300 rows from the original dataset to provide a broad range of samples for users to analyze. We then backed up the system and exported it to other team members' computers for testing, which was successful.

Sample Data

To ensure our Motor Vehicle Collisions database offers valuable insights into traffic incidents in New York City, we're carefully sourcing our dataset from a CSV file in ELMS, which is connected to the NYC Open Data platform. Our dataset comprises over 2 million records detailing various aspects of motor vehicle collisions, aiming to provide comprehensive analyses to enhance road safety.

To refine our approach and ensure our sample is manageable and representative, we plan to use a stratified sampling method. Firstly, we'll include incidents from various boroughs across New York City to achieve geographical diversity, enabling us to thoroughly examine geographical differences in collision rates and contributing factors. Secondly, our sample will cover data from different years and months to identify patterns and assess the impact of temporal shifts, reflecting seasonal variations and long-term trends. Lastly, we'll ensure our sample encompasses a range of incident severities, from minor fender-benders to fatal crashes. This will allow us to explore the complex nature of traffic collisions, analyzing the correlation between incident severity and other important variables and gaining a nuanced understanding of road safety challenges in New York City.

Snippet of Vehicle_Table showing necessary information about Vehicles

INST 327 (0103)

FINAL REPORT - 05/11/2024

GROUP: 0103-G1

GROUP MEMBERS: Gabriel Gonzalez, Weizhen Ye, Johnathan Hill, Mohammad Ibrahim, Motunrayo Aduloju

Unique_ID	Vehicle_Make	Vehicle_Model	Vehicle_Year	State_Registration	Vehicle_type	Travel_Direction	Vehicle_Occupants	Pre_Crash	Point_of_Impact
16938759	FRHT	TRUCK TRADE	2011	MD	Pick-up Truck	East	1	Going Straight Ahead	No Damage
16938968	chrysler	town and country	2015	NY	Station Wagon/Sport Utility Vehicle	Unknown	0	Parked	Left Rear Bumper
16952374	TOYT -CAR/SUV	TOYT 4RN	2015	NY	Station Wagon/Sport Utility Vehicle	South	1	Passing	Right Front Quarter Panel
16975598	isuza	deliv	2012	NY	truck	North	1	Going Straight Ahead	Left Side Doors
17044144	DODG -CAR/SUV	DODG CHA	2015	MA	4 dr sedan	East	1	Parked	Left Side Doors
17044639	MERZ -CAR/SUV		2015	NY	4 dr sedan	East	2	Merging	Right Front Bumper
17045096	FORD -CAR/SUV	FORD ZZZ	2000	NY	Station Wagon/Sport Utility Vehicle	West	1	Going Straight Ahead	Center Front End
17049525	TOYT -CAR/SUV		2014	NY	Sedan	North	0	Going Straight Ahead	Center Front End
17072438	YAMA -MCL		2012	NY	Motorcycle	North	1	Going Straight Ahead	Center Front End
17154674	NISS -CAR/SUV		2012	NY	Station Wagon/Sport Utility Vehicle	South	0	Parked	Right Rear Bumper
17172871	KIA -CAR/SUV		2016	MA	Sedan	Unknown	0	Parked	Right Rear Quarter Panel
17179435	FRHT-TRUCK/BUS		2015	NY	Flat Bed	Southeast	1	Making Right Turn	Trailer
17187987	BMW -CAR/SUV		2014	NJ	4 dr sedan	East	1	Going Straight Ahead	Center Back End
17199379	MERZ -CAR/SUV		2005	NJ	Station Wagon/Sport Utility Vehicle	Unknown	1	Parked	Left Rear Bumper
17222837	freig		2006	NY	Pick-up Truck	South	0	Going Straight Ahead	Center Back End

View/Queries

In our vehicle database project, we have developed five queries saved as views, each aimed at extracting valuable insights from our data. These queries address various aspects of our database, covering topics such as vehicle make frequencies and correlations between vehicle age and accident rates. Some of the original queries outlined in our initial proposal had to be revised or replaced due to limitations in SQL capabilities or changes in the database structure. We replaced them with new, more detailed queries that focus on tables included in the database, ensuring they meet our project requirements.

By saving our queries as views, we enhance accessibility for potential users of the database. Views provide a convenient way for users to access pre-defined query results without needing to understand or execute the underlying SQL code. This approach streamlines data access and analysis, making it easier for users to derive meaningful insights from the database.

View Name	Req. A	Req.B	Req. C	Req.D	Req. E
Query_1		X			
Query_2	X	X			
Query_3	X	X	X		
Query_4	X	X	X		X

Query_5	X			X	

The following list of Queries answers the next questions:

Query_1 (VIEW MostFrequentCollisions): "Which vehicle makes are most frequently involved in collisions?"

Query_2 (VIEW VehicleAgeCollisionCorrelation): "Is there a correlation between the age of a vehicle and its involvement in collisions?"

Query_3 (VIEW TopVehicleMakesAvgOccupants): "What are the top vehicle makes involved in collisions, and what is the average number of vehicle occupants for each make?"

Query_4 (VIEW DriversLicenseJurisdictionAccidents): "What driver license jurisdiction is involved in most accidents?"

Query_5 (VIEW AvgDamagesPerVehicleYear): "What is the average number of damages per vehicle year?"

Changes From Original Design

During the early stages of the project, our team struggled to meet expectations due to the extensive nature of our database. Establishing relationships between entities and attributes was challenging. However, feedback from our instructor and progression through different project stages clarified the expectations and methods required. We redefined relationships between entities like collisions, vehicles, drivers, and contributing factors. During normalization, we used feedback to implement changes, reducing redundancy and improving data integrity. For example, we removed the collisions table and incorporated its columns into the Vehicle table to eliminate repeated data. We also removed the Vehicle_ID column and the Vehicle-Driver Linking Table, recognizing it was a many-to-one relationship.

Furthermore, we included a linking table between the Vehicle and Damage tables called Damage_Table_Has_Vehicle to establish a many-to-many relationship. Another significant change was specifying the data type for each attribute in the tables, addressing issues from the logical design assignment. We also redesigned sample data rows to ensure they were not primarily null values. Finally, we redefined most columns as Not Null to help prevent normalization problems.

Database Ethics

During the development of the Motor Vehicle Crash Database, we prioritized key ethical considerations such as data privacy, fair use, diversity, equity, and inclusion. The dataset contains no personal information and has been approved by our professors and teaching assistants.

We ensured our sample data represented New York City's diverse population without disclosing sensitive information like race, gender, or age, reducing potential biases. Access to the system is restricted to course faculty members, minimizing the risk of data leaks.

By focusing on diversity, equity, inclusion, data privacy, and fair use, we have created a database that ethically analyzes motor vehicle crashes in New York City and reflects our understanding of data ethics.

Lesson Learned

Throughout the project, our group learned valuable lessons, including the importance of effective communication and collaboration. These skills are crucial for completing a project efficiently and with high quality. As we progressed, we better understood each other's skill sets and utilized them more effectively. We also learned that defining responsibilities and roles early on significantly enhances efficiency. When each person has a specific role, it prevents task overlap and increases focus, ultimately improving overall productivity. With many project checkpoint assignments throughout the semester, having a clear timeline would have improved our preparation and allowed us to complete tasks earlier. Additionally, maintaining a flexible mindset was crucial. Being adaptable helped us quickly resolve issues and find the best solutions

Potential Future Work

Our Motor Vehicle Collisions database provides insights into NYC traffic incidents, but expanding its scope will greatly improve its effectiveness and data storytelling abilities. By including recent and historical data and additional sources like weather and traffic volume, it will allow for a more comprehensive analysis. Adding socio-demographic information will offer a nuanced view of how different groups are affected. Additionally by adding enhanced data analysis techniques (such as predictive modeling and spatial analysis) it would better highlight trends and hotspots. By ensuring data accessibility for stakeholders and implementing feedback mechanisms will boost community engagement. Regularly updating data privacy policies and addressing biases will maintain the database's integrity. These enhancements will significantly increase its utility and impact on improving road safety.

Citations

McKinsey & Company. “What Is Diversity, Equity, and Inclusion (DE&I)? .” *Www.mckinsey.com*, McKinsey & Company, 17 Aug. 2022, www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-diversity-equity-and-inclusion.

NYPD, Police Department. “Motor Vehicle Collisions - Crashes: NYC Open Data.” Motor Vehicle Collisions - Crashes | NYC Open Data, 23 Feb. 2024, data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data.

“Research Ethics.” National Institutes of Health, U.S. Department of Health and Human Services, oir.nih.gov/sourcebook/ethical-conduct/research-ethics. Accessed 25 Feb. 2024.