INST 327, Section: 0201

Project Report

05/03/2024

Group 7: Andrew Ho, Lennon Brosoto, Arvin Singh, Stephen Thang, Kirk Laryea

Introduction:

To enhance traffic safety, the NYPD gathers and oversees data concerning incidents tied to traffic, such as motor vehicle collisions. A motor vehicle collision in NYC since July 1st, 2012 are represented by each row in the dataset. A collision is mandated if it involves injuries, fatalities, or property damage exceeding $1000.

We will be collecting, analyzing and managing data on the total number of SUV collisions, car model most likely to get in a collision, number of injuries, drivers gender(male), the year in which the collision occurred, number of drivers without license registration, date of collision, amount of collisions on major holidays, the vehicle make with the recorded most amount of collisions, the number of collisions resulting in the damaging of property, the state of the vehicle upon collision and the time of the day the collision occurred from New years in years 2012- present. The primary objective of our project is to gather, analyze, and interpret data on vehicle collisions in New York. Our objective is to help improve traffic safety by finding its causes and reducing them.

We chose this topic because motor vehicle collisions are becoming a daily occurrence in NYC. By examining and understanding the causes of collisions, we aim to inform drivers, policymakers, law enforcement agencies, urban planners and traffic engineers in enforcing effective road measures to enhance road safety, reduce the number of deaths and promote a culture of road awareness and safe driving throughout New York.

INST 327, Section: 0201

Project Report

05/03/2024

Group 7: Andrew Ho, Lennon Brosoto, Arvin Singh, Stephen Thang, Kirk Laryea

Database Description:

Our database will focus on the vehicle collisions that have occurred in New York City, on New Year's Eve the year of 2017. We've decided to narrow down the Motor Vehicle Collisions dataset to this state and day because we agreed that the state of New York, and specifically New York City is a culturally diverse area, which ensures that our data collection is inclusive. We decided to analyze motor vehicle collisions on New Year's Eve because of it being a universally celebrated holiday as well as the increased levels of traffic. Our plan for the database is to present the data in nine tables that correlate with each other that presents information such as the vehicle info, driver info, and collision info. There will be other information that pertains to the damage of the vehicle, the contributing factors, and collision and property damage information. All of this will be done by organizing our database using the "unique_id" as our primary key, in the vehicle info table. Through this primary key, we can connect other information that pertains to a vehicles collision by connecting them to foreign keys such as the "collision_id", "vehicle_damage_id", "contributing_factor_id", "driver_id", "public_property_id", and the "public_property_damage_id".

INST 327, Section: 0201

Project Report

05/03/2024

Group 7: Andrew Ho, Lennon Brosoto, Arvin Singh, Stephen Thang, Kirk Laryea

Views/Queries:

View Name	Req A	Req B	Req C	Req D	Req E
type_car_most_likely_in_accident	**X**	**X**	**X**		
men_vs_women_sedan_collision	**X**	**X**	**X**		
vehicle_type_and_public_property_damage	**X**	**X**		**X**	
car_company_most_collisions		**X**	**X**		**X**
most_collisions_time_period	**X**		**X**		

We created five queries saved as views for our database. These queries answer questions that can be asked with the data from our database such as, gender correlation to vehicle collisions, the vehicle type or company that has experienced the most collisions, public property damage, and time correlation to collisions. These queries answer many questions that were asked in our original proposal, some of which were revised or added to the questionnaire. We decided to save these queries as views so the user has easy access to this information.

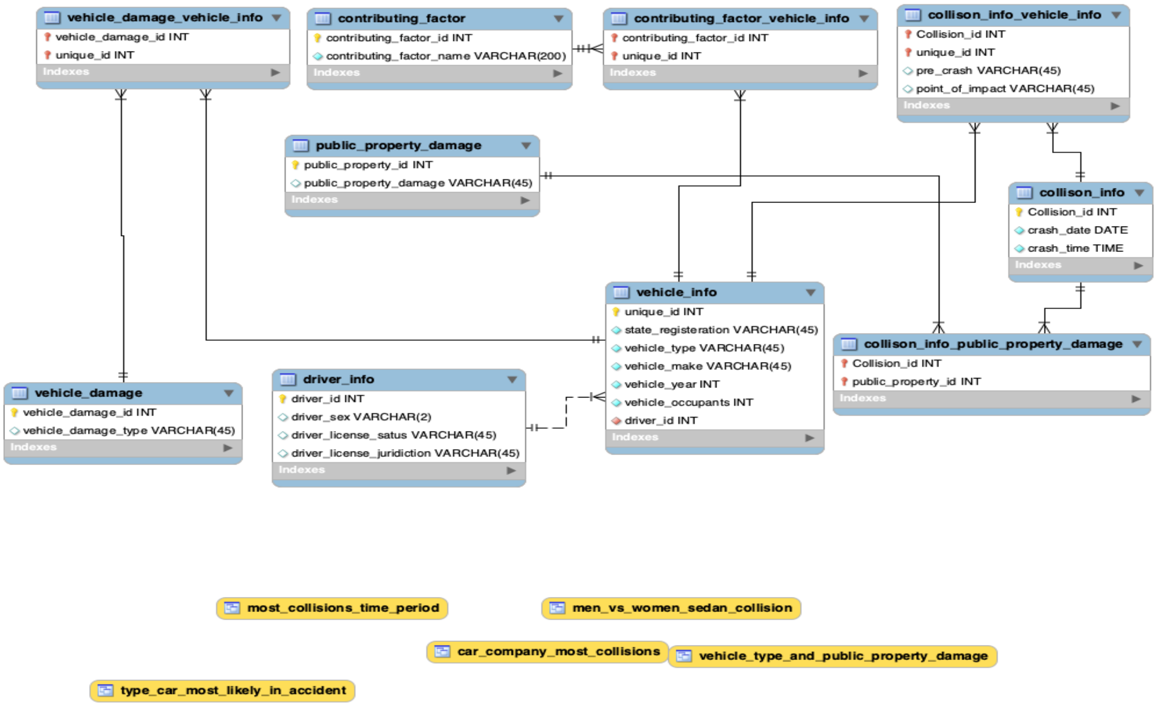
Use of * Explanation:

The use of the wildcard character (*) in our view/query called "type_car_most_likely_in_accident" was unavoidable given the fact of the nature of the query itself. The query itself needs to access all of the data in order to give a proper representation Of what type of cars get into the most amount of accidents. In this case and scenario it was unavoidable and was necessary in order to develop and create the query in order to get the proper necessary results.

Physical Design:

Our database includes ten tables. Our main table is the vehicle_info table and it has four reference tables like, driver_info, vehicle_damage_vehicle_info, and contributing_factor_vehicle_info, collision_info_vehicle_info. In our database, our tables signify the relationships between driver, vehicle, and collision. For example, each collision can have multiple vehicles involved, a one to many relationship between collision and vehicle. Each collision can have multiple drivers, a one to many relationship between collision and drivers. And a driver can have multiple vehicles, one to many between drivers and vehicles.

Logical Design:


vehicle_damage_vehicle_info
vehicle_damage_id INT
unique_id INT
Indexes
contributing_factor
contributing_factor_id INT
contributing_factor_name VARCHAR(200)
Indexes
contributing_factor_vehicle_info
contributing_factor_id INT
unique_id INT
Indexes
collison_info_vehicle_info
Collision_id INT
unique_id INT
pre_crash VARCHAR(45)
point_of_impact VARCHAR(45)
Indexes
public_property_damage
public_property_id INT
public_property_damage VARCHAR(45)
Indexes
collison_info
Collision_id INT
crash_date DATE
crash_time TIME
Indexes
vehicle_info
unique_id INT
state_registeration VARCHAR(45)
vehicle_type VARCHAR(45)
vehicle_make VARCHAR(45)
vehicle_year INT
vehicle_occupants INT
driver_id INT
Indexes
collison_info_public_property_damage
Collision_id INT
public_property_id INT
Indexes
vehicle_damage
vehicle_damage_id INT
vehicle_damage_type VARCHAR(45)
Indexes
driver_info
driver_id INT
driver_sex VARCHAR(2)
driver_license_satus VARCHAR(45)
driver_license_juridiction VARCHAR(45)
Indexes
most_collisions_time_period
men_vs_women_sedan_collision
car_company_most_collisions
vehicle_type_and_public_property_damage
type_car_most_likely_in_accident

INST 327, Section: 0201

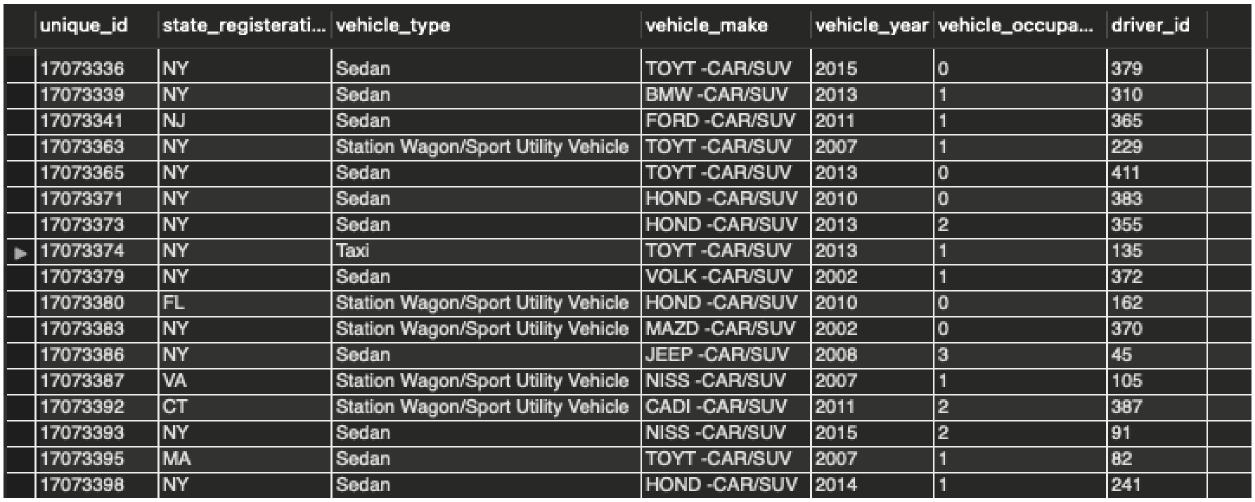
Project Report

05/03/2024

Group 7: Andrew Ho, Lennon Brosoto, Arvin Singh, Stephen Thang, Kirk Laryea

For our database design, we tried to incorporate data that would showcase the necessary information to conclude possible reasons for a vehicle collision. We wanted to share information relevant to vehicle collisions such as information about the driver, information about the vehicle, and information pertaining to the vehicle crash such as property damage, time, pre-crash and point of impact information. We normalized our database so that the data represents its relationships with other data in the database.

Sample Data:



	unique_id	state_registerati...	vehicle_type	vehicle_make	vehicle_year	vehicle_occupa...	driver_id
	17073336	NY	Sedan	TOYT -CAR/SUV	2015	0	379
	17073339	NY	Sedan	BMW -CAR/SUV	2013	1	310
	17073341	NJ	Sedan	FORD -CAR/SUV	2011	1	365
	17073363	NY	Station Wagon/Sport Utility Vehicle	TOYT -CAR/SUV	2007	1	229
	17073365	NY	Sedan	TOYT -CAR/SUV	2013	0	411
	17073371	NY	Sedan	HOND -CAR/SUV	2010	0	383
	17073373	NY	Sedan	HOND -CAR/SUV	2013	2	355
▶	17073374	NY	Taxi	TOYT -CAR/SUV	2013	1	135
	17073379	NY	Sedan	VOLK -CAR/SUV	2002	1	372
	17073380	FL	Station Wagon/Sport Utility Vehicle	HOND -CAR/SUV	2010	0	162
	17073383	NY	Station Wagon/Sport Utility Vehicle	MAZD -CAR/SUV	2002	0	370
	17073386	NY	Sedan	JEEP -CAR/SUV	2008	3	45
	17073387	VA	Station Wagon/Sport Utility Vehicle	NISS -CAR/SUV	2007	1	105
	17073392	CT	Station Wagon/Sport Utility Vehicle	CADI -CAR/SUV	2011	2	387
	17073393	NY	Sedan	NISS -CAR/SUV	2015	2	91
	17073395	MA	Sedan	TOYT -CAR/SUV	2007	1	82
	17073398	NY	Sedan	HOND -CAR/SUV	2014	1	241

This data includes information about the vehicle identification, state registration, type of vehicle, make, year, occupants and the drivers_id.

Changes From Original Design:

In terms of changes from our original design we made a lot of changes with the linking tables and some foreign keys as well. Below are the changes we made;

1.Vehicle_damage_driver to vehicle_damage_vehicle_info(linking table): We realized that this would be more relevant to link vehicle damage with vehicle info since the driver is already linked with the vehicle_info table.

2. We changed vehicle_id in the vehicle info table to unique_id due to vehicle_id repeating for some rows.

3. We created a new table linking property damages and collision info.

4. We created a new linking table for collision_info and vehicle_info.

5. We moved pre_crash and point_of_impact to the collision_info_vehicle_info linking tables because they vary on both on vehicle and collision.

6. We created a new linking table between contributing_factor and vehicle_info.

7.We linked the driver_info table to vehicle_info table by making driver_id a foreign key.

These are the changes we've made so far and after reviewing with our mentor, everything seems to be on the right track.

Database Ethics Considerations:

Our viewpoint about the ethical considerations for using open-source data remains the same as in our initial proposal. The data used was released to the public, is updated frequently, and allows downloading and using it freely. It shouldn't pose any potential violations of privacy, copyright, or other legal concerns due to the free access to information. Sources knowledgeable in data ethics define public data as being shared, used, reused, and redistributed without restriction (Hashemi-Pour, 2024); our use of publicly available data fits the description and allows us to use the data without limitation. In the discussion about privacy, our use of public data does not violate privacy laws since the information provided within the database doesn't reveal sensitive, personally identifiable information about someone, is ethically sourced and released to the public, and wasn't stolen or breached, all of which is considered ethical use of open data (University of Wolverhampton). We have also been focused on ensuring inclusivity and fairness in our database, aiming to capture the full range of social, historical, and demographic diversity within our database. We are able to demonstrate diversity in our data through our collection of collision data from all vehicle types, regardless of gender, as well as recording data from New Year's day, which is a global holiday to represent all social demographics. We also decided that New York City holds a population of various cultures and demographics, focusing our data collection from that area. Our data collection is not sorted on pre-crash scenarios and contributing factors to ensure that we are collecting vehicle collisions from all sorts of events.

Lessons Learned:

During the creation of this final team project, we've faced many challenges ranging from technical issues, internal communication, and planning and structuring of our database. Upon choosing the Motor Vehicle Collision database, we realized that the CSV itself contained too much information for it to be useful within our database, therefore it became apparent that we would have to limit our data insertion to information that pertained to New York City on January 1st. We chose to narrow our data to January 1st in New York City because of our newfound influence of data ethics, diversity, and inclusion. We decided to choose New York City and January 1st because New Years being a universally celebrated holiday by all people, and that New York City had a very diverse population of different backgrounds and ethnicities.

Potential Future Work:

Looking into the future, there are multiple ways we could expand this project. Firstly, our project captures data only from New Year's day in New York City in 2017. We could focus our database on New Year's day vehicle collisions, but offering the collision data from multiple years. By adding more information from multiple New Years, we could determine potential trends and patterns in vehicle collisions. Another way we could expand on our database would be to include information such as injury severity of passengers, where we can see trends in injury to types of crash scenarios. In the future, we may also expand the data to include other major holidays of the year, to identify any unique characteristics specific to each holiday.

References

Hashemi-Pour, Cameron. “What Is Public Data? | Definition from TechTarget.” *CIO*, June 2023, www.techtarget.com/searchcio/definition/public-data#:~:text=Public%20data%20is%20information%20that%20can%20be%20shared%2C. Accessed 23 Feb. 2024.

NYC OpenData. “Motor Vehicle Collisions - Crashes | NYC Open Data.” *Data.cityofnewyork.us*, 7 May 2014, data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data.

University of Wolverhampton. “Ethics & Open Data - University of Wolverhampton.” *Www.wlv.ac.uk*, www.wlv.ac.uk/research/research-policies-procedures--guidelines/ethics-guidance/ethics--open-data/#:~:text=In%20research%20 with%20 people%20there%20 may%20be. Accessed 18 Apr. 2024.