Analysis of Accidents in New York

Concepts of Database Systems Project

Dr. Fang Jin

Spring Semester 2019

Anish Basnet
Bhimsen Bhattarai
Celina Maharjan
Rajeev Kuikel

1. Introduction

The document is a detailed report on the Concepts of Database Systems Project by “Team 6”. The Team has a total of 4 members: Anish Basnet, Bhimsen Bhattarai, Celina Maharjan, and Rajeev Kuikel.

From the pool of the database containing thousands of un-organized information on the accidents that occured in New York, the main objective of the team was to be able to provide user-specific data. By user-specific data, we mean, information such as: the action of the driver right before the occurrence of the accident, the direction the vehicles were heading towards during the accident, the number of specific make and model of the vehicles involved in accidents, and the number of specific vehicle body type involved in the accident during the timeline in New York according to the data set provided.

2. Datasets

The datasets that we chose for this project are extracted from source 1: “Motor Vehicle Crashes in New York.” The dataset that we have used for this project is limited by time and by the number of data selected for the project. The dataset contains the accidents that were reported till the year 2016.We divided the total database into two tables.

One of the tables contains the vehicle information while the other contains the contributing factors such as: direction of the vehicle at the time of crash, action of the driver, what caused the accident, and so on.

	A	B	C	D	E
1	Accident ID	Vehicle ID	Action Prior to Accident	Direction of Travel	Contributing Factor
2	5000		Going Straight Ahead	East	HUMAN
3	5001		Merging	South	HUMAN
4	5002		Going Straight Ahead	South	HUMAN
5	5003		Going Straight Ahead	North	ENVMT

Figure 2.1: Accident-table.csv

The diagram above is the snip of the accident-table.csv. The accident table, like mentioned above has been exported from the data that we accessed from source 1. This table contains information such as the ‘action prior to accident’, ‘direction of travel during accident’, and ‘contributing factor: human or environment.’

	A	B	C	D
1	VEHICLE ID	Vehicle Body Type	Vehicle Year	Vehicle Make
2	1	SUBURBAN	2015	NISSAN
3	2	4 DOOR SEDAN	2007	HYUNDAI
4	3	4 DOOR SEDAN	2009	HONDA
5	4	4 DOOR SEDAN	2007	DODGE
6	5	SUBURBAN	2012	MITSUBISHI

Figure 2.2: Vehicle-table.csv

The figure above is a snip of the Vehicle-table.csv which is also a excerpt of the data found in source 1. This data table only contains information on the Vehicle. As can be seen on Figure 2.2, the data table contains the ‘Vehicle body type’, ‘Vehicle Year’, and ‘Vehicle Make’.

3. Database Design

3.1 Conceptual Database Design

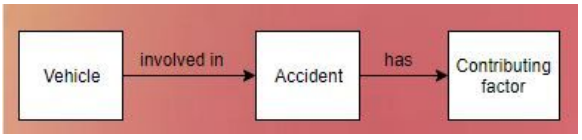

Figure 3.1.a: Conceptual Design

The diagram above, is the conceptual design of the “Accident-table” dataset mentioned in section 2. There are three

entities in *Figure 3.1.a*: Vehicle, Accident, and Contributing factor. The first two entities, Vehicle and Accident are connected with a 'involved in' relationship. Each vehicle has been involved in one or more accidents based on our dataset.
The last two entities, Accident and contributing factor is connected with a 'has' relationship. Each accident has one or more contributing factors.

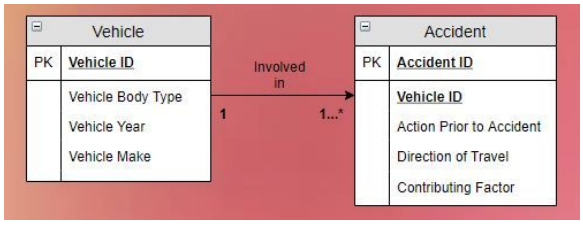

Figure 3.1.b: Entity Relation Diagram

The figure above, is the conceptual design of the "Vehicle-table" dataset which has also been mentioned in section 2. There are two entities for this dataset: Vehicle and accident. The two entities are connected with a 'involved in' relationship. Each vehicle in the dataset has been involved in one or more accidents.

3.2 Logical Database Design

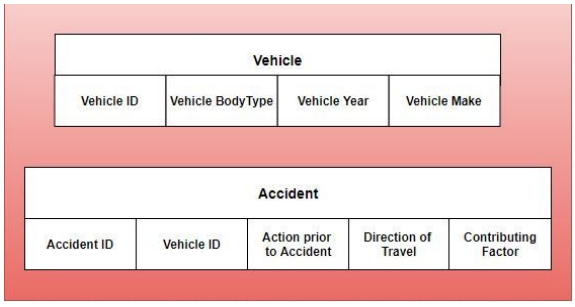

Figure 3.2: Conversion of ER diagram with two entities into two relational tables

In the figure above, the relationship between two tables have been reflected. Both the 'vehicle' and 'accident' tables have an attribute 'VEHICLE ID'. The 'VEHICLE ID' attribute in the vehicle table contains the identification number for each vehicle which is unique and the foreign key for the table accident table. We join the two tables based on this attribute.
The table is the normalized form of the original dataset that we pulled from Source 1 as there were many dependencies in the original dataset.

3.3 Physical Database Design

Figure 3.3: physical Database Design

The *Figure 3.3* above shows the physical database design for the original table. The table contains accident id, vehicle id, action prior to accident, direction of travel, and contributing factor.

3.4 Index and Constraints

3.4.1 Index

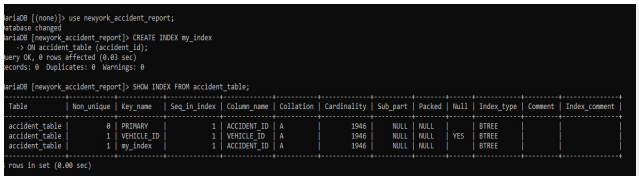

Figure 3.4.a: Index Diagram

The figure above is shows the Create Index code. The attribute that is mostly used is the accident_id. We index the attribute

mentioned above by using Create Index statement shown in the diagram above.

3.4.2 Constraints

```
MariaDB [newyork_accident_report]> ALTER TABLE accident_table
    -> ADD FOREIGN KEY (vehicle_id)
    -> REFERENCES vehicle_table(vehicle_id);
Query OK, 1946 rows affected (0.10 sec)
Records: 1946  Duplicates: 0  Warnings: 0

MariaDB [newyork_accident_report]>
```

Figure 3.4.2 a: Foreign Key

We added a foreign key constraints on the accident table which references the vehicle id attribute of the vehicle table.

```
| 804 | TRUCK           |          2003 | GMC           |
| 812 | TRUCK           |          2009 | HONDA         |
| 813 | SUBURBAN        |          2005 | GMC           |
| 814 | 4 DOOR SEDAN    |          1999 | BMW           |
| 815 | 4 DOOR SEDAN    |          2002 | FORD          |
| 833 | SUBURBAN        |          2012 | NISSAN        |
+-----+-----------------+---------------+---------------+
17 rows in set (0.01 sec)

MariaDB [newyork_accident_report]> ALTER TABLE accident_table
    -> ADD PRIMARY KEY(accident_id);
```

Figure 3.4.2 b: Primary Key

We added a primary key constraint on the accident table which references the accident ID attribute of the accident table.

4. Implementation

4.1 Validation

For the validation and testing of our project, we performed cross validation by inserting a row, deleting a row and updating data in the existing rows in the dataset.

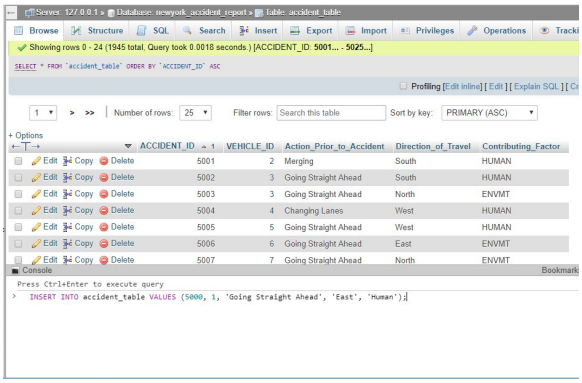

Figure 4.1.a: Insert Validation

In the figure above, we have executed the insert sql query. For validation, we tried to insert a new row with id ‘5000’.

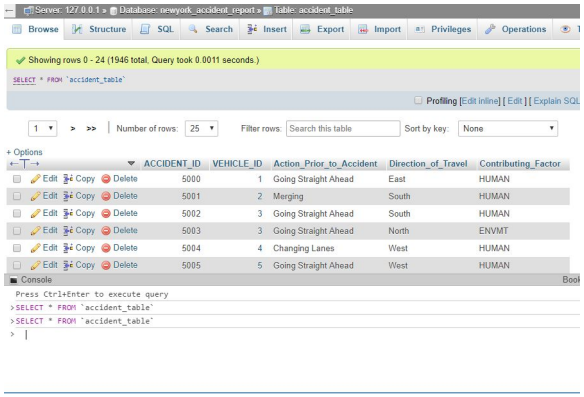

Figure 4.1.b: Insert Validation Result

The figure above represents the result of the ‘insert’ query. As it can be seen, the first row is affected and thus a new row with id ‘5000’ is inserted in the table.

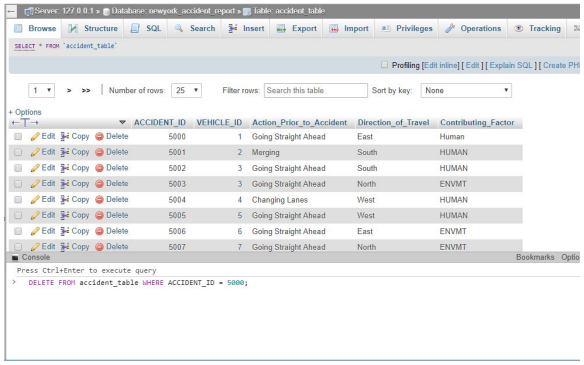

Figure 4.1.c: Delete Validation

In the figure above, we have attempted to execute the ‘delete’ query for validation. We are deleting the recently inserted row with id ‘5000’.

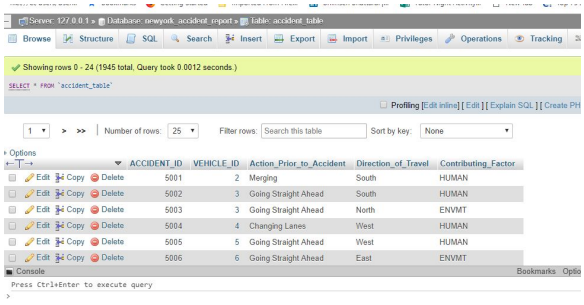

Figure 4.1.d: Delete Validation Result.

The figure shows the result of the delete query. As it can be seen in the table above, the newly inserted row with ID ‘5000’ is deleted as a result.

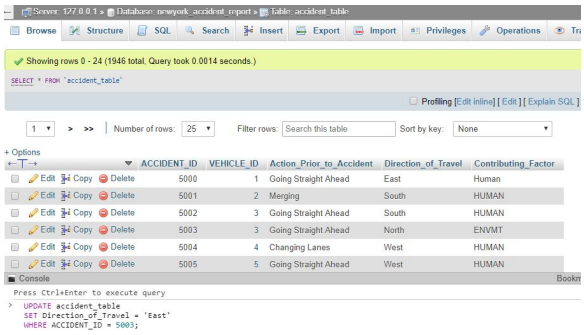

Figure 4.1.e: Update Validation

In the figure above, we have executed the Update sql query. For validation, we tried to update the ‘direction of travel’ for rows with vehicle id ‘5003’.

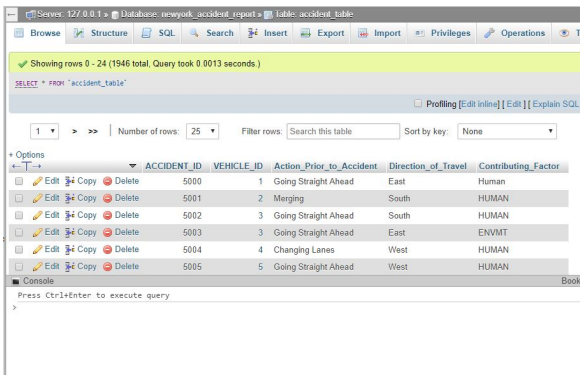

Figure 4.1.f: Update Validation Result

The figure above shows the result of the update query executed in *Figure 4.1.e*. As you can see, the direction of travel for vehicle ID 5003 is updated from ‘North’ to ‘East’.

5. Conclusion

5.1 Challenges

The datasets we used were raw and in unorganized form. There were several errors in some of the data table. Some text were put in error to different categories. For this, we used MS Excel find and replace tool to fix that error. In the direction attribute had Truck in the list which is not possible as there are 4 directions. This was our first challenge which made some difficulties in data visualization. Another challenge we faced was to link two table during visualization.

5.2 Contribution

Our Project’s two main goals were to find out the contributing factor behind the accident and prior action which led to the accident. We believe those two would be the major finding from this project. Several other goals were to see if accident is higher according to the type of vehicle body type and year. We were concerned to see if vehicle body type and year would even be a factor of accident cause. We were thrilled to see that trucks were involved in accident more than other vehicle like SUV and Sedan Cars (Fig: 5.2.f).

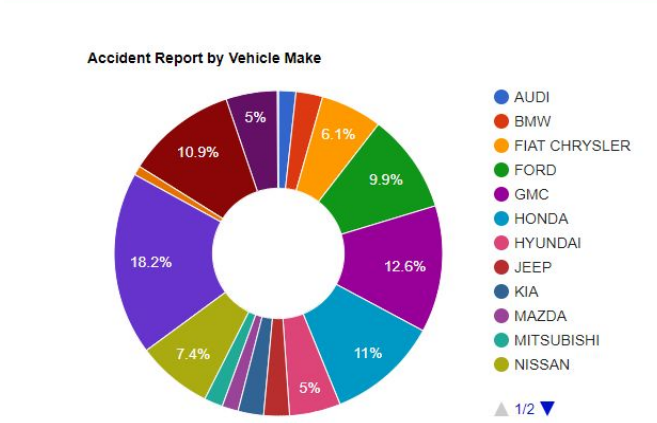

Fig: 5.2.a: Diagram showing accident report by Vehicle Make

We cannot really determine if vehicle make may be the reason behind any accident. Our data had many leading car companies like Honda, BMW, Jeep, Mazda etc. The one thing that could be pulled up from the dataset we had was the expensive cars brand like BMW and Audi were involved in higher number of accidents compared to other car brands. Our dataset only consists of accident reports from New York so we think we can not make any conclusion from this report we queried which displays out the percentage of accident according to the Vehicle make(Fig 5.2.a).

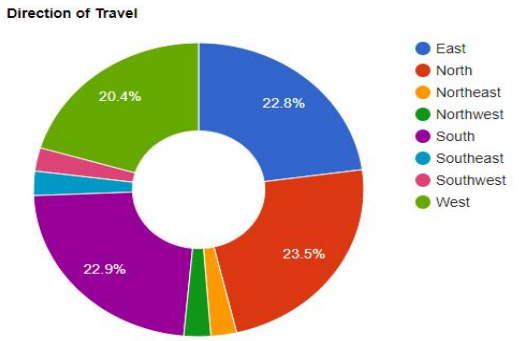

Fig: 5.2.b: Diagram showing accident report according to vehicle's direction of travel

It was interesting to see that Vehicle travelling all four main direction had almost equal number of accidents. It is certain that direction of travel does not contribute to the cause behind accident(Fig: 5.2.b).

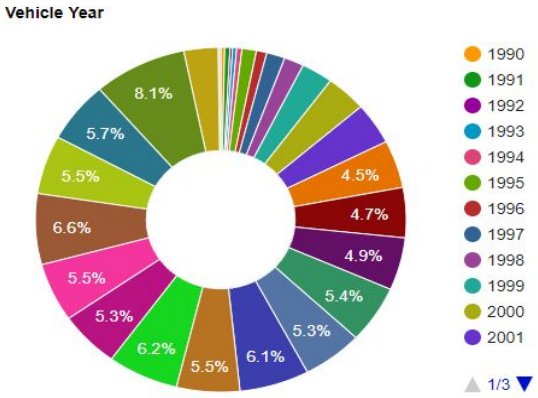

Fig: 5.2.c : Report by Vehicle Year

Like we were not able to determine if vehicle may be the reason behind accident, similarly, we are certain that vehicle year cannot be the cause of the accident. As seen in Fig: 5.2.c, it is clear that vehicle year can not be the factor of accident because from the datasets we had there were not significant number of vehicle of only specific year involved in accidents.

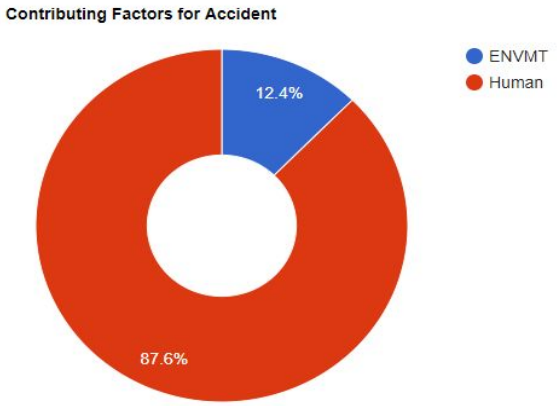

Fig: 5.2.d: Diagram showing two main contributing Factors for accident

One of the major finding from the datasets we had was to determine the responsible factor behind the accident. 87.6% accidents were caused by the human factor and only 12.4% accidents were caused by any environmental factor(Fig: 5.2.d). It was not clear what specific human factors could be responsible for that many accidents but we can surely say that almost all accidents are the sad consequence of human negligence.

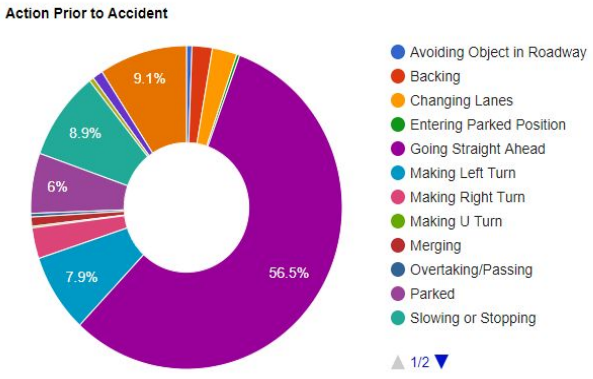

Fig: 5.2.e : Diagram showing Report by Action Prior to accident

From the above figure Fig: 5.2.d, we could draw the conclusion that the main factor behind the accident is human negligence and also Fig: 5.2.e evidently underpins that human beings play paramount role to result in any accident. There can not be any other reason why vehicle heading straight could have the highest number of accidents but human carelessness. Fig 5.2.e expose the several actions prior to the accident from which it is understandable that going straight ahead vehicle got into accident 56.5% more than any other act that one could sense of. Many people would think that changing lanes could impact on accidents but from the result changing lanes contributed only 9.1%.

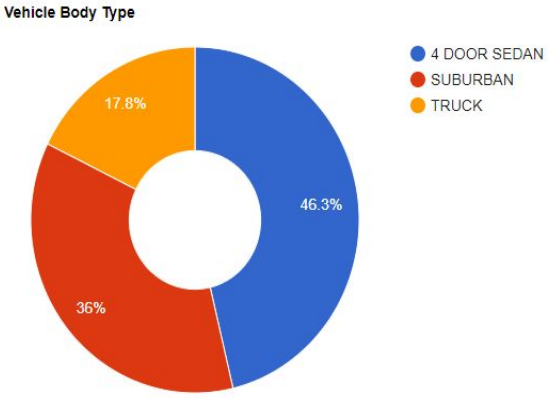

Fig: 5.2.f: Diagram showing Accidents by three common kinds of Vehicle body type

6. Acknowledgement

Anish: Visualization, Constraints.
Bhimsen: SQL Queries, Data Cleaning, Indexes, and Constraints.
Celina: Conceptual, Logical, and Physical Model design.
Rajeev: Cross-Validation, GitHub.

7. Tools Used

1. Draw.io: Draw.io was used for the design and completion of most of the diagrams present in this documentation.
2. Microsoft Excel: Microsoft Excel was used to access/filter the data file which was stored as a .csv file.
3. Version Control: GitHub was utilized as version control to upload the code and the report of the project.
4.Google Chart API: This API was used to create interactive charts for data visualization.
5. Softwares: HTML, CSS, PHP
6. Xampp: Xampp was used as web server solution.

8. Project Github Link

https://github.com/RajeevKuikel/DataBase-Project

9. Sources

[1] https://catalog.data.gov/dataset/motor-vehicle-crashes-vehicle-information-beginning-2009