UNIVERSIDAD
DE MÁLAGA

Master en Advanced Analytics on Big Data

Knowledge Area: Big Data

Task 6: Data search and visualization

Author: Juan Morales Conde

Juan Rafael Caro Moreno

Martín Blázquez Moreno

Module: Data Analytics

Málaga, 12^{th} January 2020

Two data sets belonging to the New York public data have been selected.

It has been assumed that the analysis will be performed for an insurance company.

The data sets that have been chosen deal with vehicle collisions along with their characteristics.

The first data set called Motor Vehicle $Collisions - Crashes$ provides a lot of information (345 MB formed by 1.63M of rows and 29 columns), the most relevant information for the analysis will be commented:

- CRASH DATE.
- BOROUGH.
- ZIPCODE.
- LOCATION.
- STREET INFORMATION.
- SUMARY PEOPLE INJURED.
- SUMARY PEOPLE KILLED.
- REASON OF THE COLLISION.
- VEHICLE TYPE.

Each row in the database represents the information of an accident.

Dataset URL:

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

The second set of data is called Motor Vehicle $Collisions - Vehicles$ also provides a lot of information (521MB consisting of 3.27M of rows and 25 columns), the most important will be cited and explained:

- VEHICLE INFORMATION
- DRIVER SEX

- DRIVER INFORMATION
- VEHICLE DAMAGE
- REASON OF THE COLLISION
- PUBLIC PROPERTY DAMAGE

Each row in the database represents the vehicle information involved in an accident.

Dataset URL:

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4

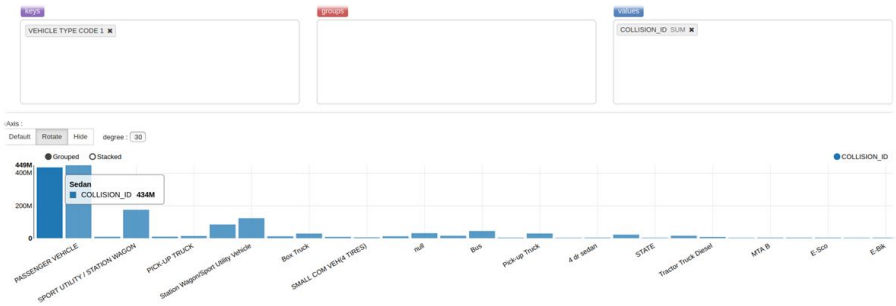
You can relate both sets of data using the $COLLISION\_ID$ column, since this column is an attribute that uniquely identifies the accident with the vehicle.

The analysis of these data sets could provide value to insurance companies, general traffic management, car brands ...

It is very interesting to carry out this analysis to improve the signaling of the streets, to ensure both pedestrians and cyclists or motorists since it provides a considerable volume of data, in addition to other sets also related to these.

A quick visualization of the $CRASHES$ database is carried out.

The following image shows the number of accidents depending on the type of vehicle.



An aggregation function has been used to visualize the number of injuries per neighborhood and the average for cyclists:

BOROUGH	Num people injured	Average Cyclist injured
0	1	3291249
null	496575	0.013633223985613625
QUEENS	303534	0.01738173599923568
BROOKLYN	353446	0.03037204696562456
BRONX	158954	0.016067440470573433
MANHATTAN	272096	0.029805328207747915
STATEN ISLAND	48961	0.006229447927942648

An SQL query has been made (generating a temporary table) to know the number of total accidents of people who drive without a license:

```
%pyspark

BD_Vehicles.registerTempTable("VEHICULOS")
```

Took 0 sec. Last updated by anonymous at January 10 2020, 2:20:27 PM.

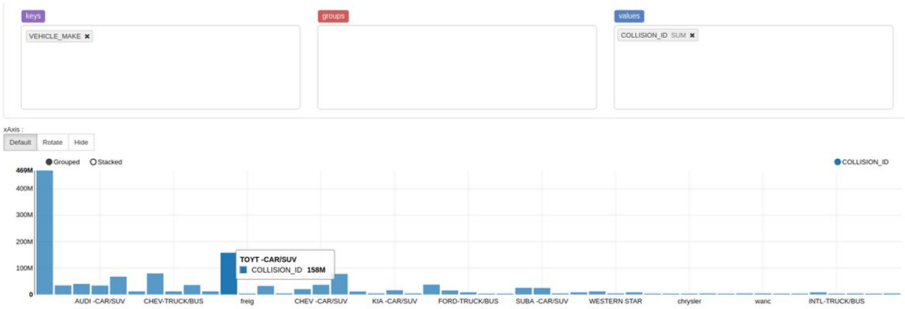
```
%sql
SELECT COUNT(*) as Number_of_Unlicensed_Driver from VEHICULOS WHERE DRIVER_LICENSE_STATUS = "Unlicensed"
```

settings

Number_of_Unlicensed_Driver
19563

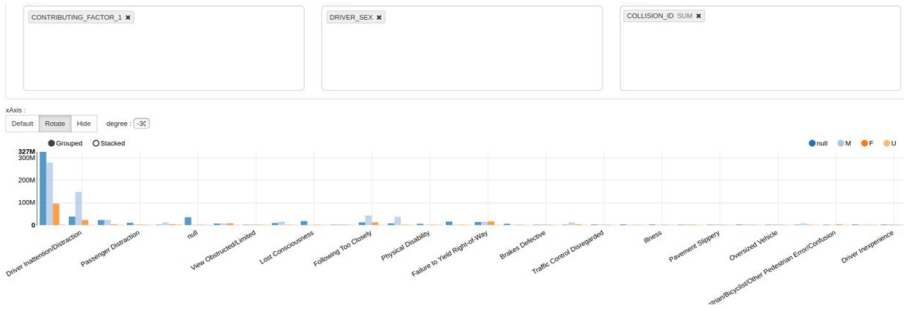
proposed questions:

1. **What are the brands of vehicles that have suffered the most accidents?**

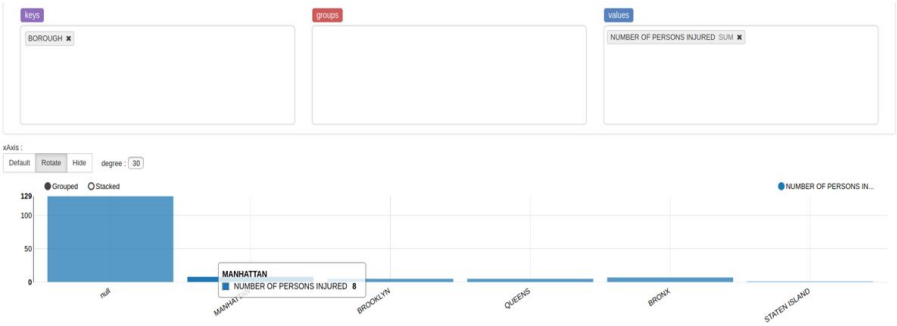


   Although some of the data is null, the answer to the question has been considered to be the TOYOTA brand.

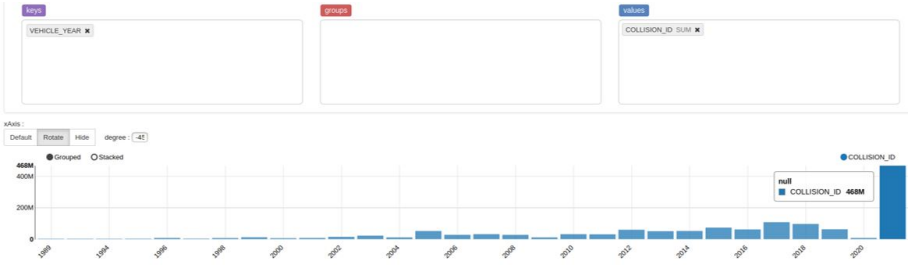
2. **Does the sex of people influence the number and cause of accidents?**



3. **What is the neighborhood where more people have been injured in an accident?**



4. **How has the number of accidents per year evolved?**



Many invalid data are also shown for this visualization (null), although an evolution of the number of accidents per year of vehicle manufacture can be observed.

This shows that, although more modern cars are better equipped with safety measures, it can make the driver more confident behind the wheel. Also, it can be considered that drivers with older cars may become more cautious with their vehicles.