NYC Vehicle Collision Report

Chan Hyun Yoo and Benno Zhang

1 Abstract

The purpose of the study was to investigate vehicle collisions in New York City based on the point of impact and different vehicle brands. Besides this objective, we also compared our results with data from the USA vehicle safety rating system in trying to answer the research problem of which car brand was the most successful at reducing injuries when there was a collision. Our design of the study consisted of two major parts: obtaining insights from NYC Car Collisions datasets from NYPD and obtaining auxiliary information from NHTSA Car Safety Ratings data. After we obtained the results from both components, we compared them to see whether the results were consistent with each other. Our results showed that the center front and center back positions were the points of impact that had the highest number of collisions. Moreover, out of all the car brands, Volvo had the lowest injury rate when encountering a collision. Similarly, from the Car Safety Ratings data, Volvo had the highest average ratings across all points of impact. We concluded that our results were consistent in showing that Volvo was the most successful brand to protect its occupants from injury during a collision in New York City.

2 Introduction

With more and more cars joining the road every year and especially with New York City being bad in traffic, the importance of finding a safe car was spotlighted. To be a safe car, we believe that it should show not only the highest performance in a car safety test but also minimum fatality when it gets into a crash on the road. Prior to this study, the range of our knowledge in this field was limited to what we see in advertisements or news presentations on car safety. According to [Cicchino, J.B. (2016)], as an example, it is shown that Volvo is excellent at reducing front-to-rear crash rates. This is not enough to fully answer the main question about the search for the best possible option for the safest car to drive in New York City.

To reach the objective of this study, we have decided to combine both the ratings given by the Department of Transportation and the fatality results from the crashes that happened in New York City. With official scores given by a trustworthy institute and the real data obtained from the roads of New York City, we are anticipating to be getting a better understanding of a reliable car brand based on safety criteria.

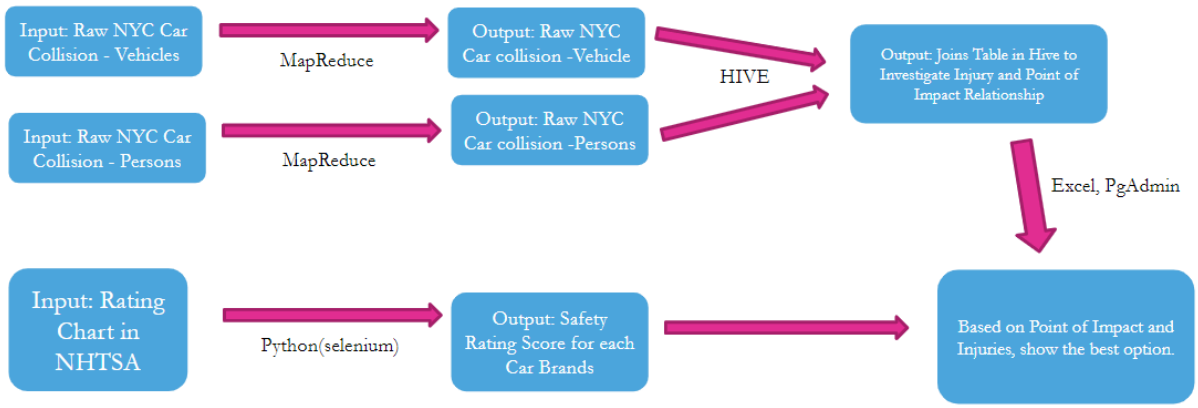


Figure 1: Diagram of Workflow

The first two inputs in the diagram were from the NYC Car Collisions dataset, which were relational tables that could be joined together via a foreign key. We used MapReduce to remove unwanted columns and rows that contained too little information. We then imported the outputs of the two tables into HIVE via HiveQL statements. Next, we used an Inner Join query to join the two tables into a single HIVE table and obtained insights regarding the relationship between occupant injury and point of impact. We used Excel and PgAdmin to verify our results on a local computer and created visualizations using Tableau.

Ratings from the Department of Transportation were also crucial data needed. Provided as a chart on the website, we decided to use Python, especially Selenium, to automate interactions on the website to scrape the data. Then, Hadoop MapReduce was used to do the statistical analysis on each brand from the CSV files we created through Python. With all data gathered for analysis, we went on looking for the best possible option for a safe car brand.

3 Motivation

With the advancement of technology, there are a variety of cars we can choose to drive and the manufacturers are doing their best to make their cars stand out. However, both car manufacturers and customers should not forget the importance of safety when choosing a car to purchase. Among design, performance, and other features, safety should be of the utmost importance in a car. We would like to expect all the passengers to get from one point to another safely.

[Chan Hyun Yoo] In the States, there were many occasions when driving a car was the only possible option to get somewhere. Frequent use of the car and especially long drives sometimes led to the thought of which car would be the best option to ensure my safety. Coming into New York City, the traffic was more tight and it seemed no surprise to see big or small accidents on the streets. With all these factors, the importance of car safety emerged as a theme to quench my curiosity.

[Benno Zhang] Personally, I have seen many Volvo commercials marketing their safety features as selling points. After further research, I learned that it was Volvo engineers who first invented the three-point seat belts that became equipped in all cars today. I wonder how Volvo, a car brand so dedicated to ensuring occupant safety, performs in real collisions and whether their statements about reducing injuries hold true from real data in New York City.

4 Related Works

We looked into research papers that conducted case studies of Volvo in trying to understand the effort Volvo has been putting into reducing occupant injury. Since we are using real vehicle collision data from New York City, we also looked into the occupant injury and fatalities history in New York City to obtain some background information for our study.

1. Safety Technology of Volvo

Numerous studies and experiments have been conducted to check and assess the safety features of Volvo.
Cicchino et al. [Cicchino (2016)] evaluated the effectiveness of forward collision warning and autonomous braking by comparing the police reported crash data with those from other car models.
Eichelberger et al. [Eichelberger (2013)] evaluated the performance of safety features of Volvo models by investigating the responses from the actual Volvo drivers.
Jakobsson et al. [Jakobsson(2010)] used the real-world accident data that involved Volvo cars with a side impact protection system and found a significant decrease in side impact related fatalities from those Volvo vehicles.

2. Motor Vehicle Fatality in New York City

Reports on motor vehicle crashes and the fatality results gave insights into how the car brands are actually showing their safety performance when out on the road.
[Department of Health and Mental Hygiene (2017)], provided insights into the characteristics of motor vehicle driver and passenger fatalities in New York City.

5 Description of Datasets

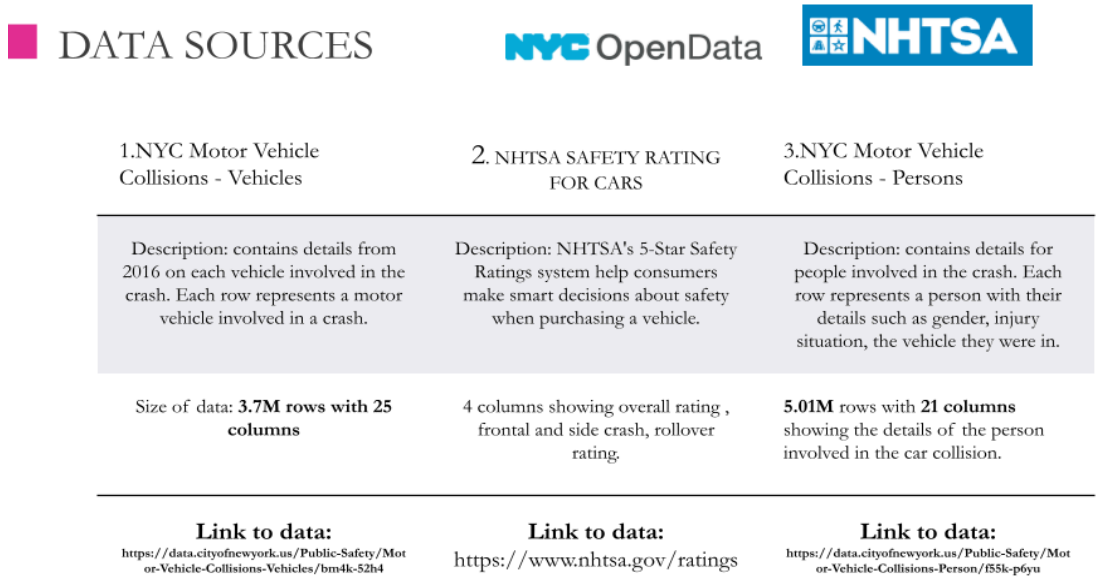


Figure 2: Diagram of datasets used and their details

Figure 2 shows the description, size, and link of the datasets used to conduct the study. Since there are too many columns in some of the datasets, we only show schema for useful columns for each dataset.

1. NYC Motor Vehicle Collisions - Vehicles from NYPD

Fieldname	Unique_id	Vehicle_Type	Vehicle_make	Vehicle_model	Point of Impact	Pre-crash
Data Type	Primary key	String, possibly categorical	String	String	String	String
Brief description	Unique record code generated by the system for each vehicle involved in a collision.	A short label of the type of car such as “sedan” or “SUV”	Brand of the vehicle, such as “BMW”	Model name of the vehicle	Location on the vehicle of the initial point of impact	Pre-crash action: Going straight, making a right turn, passing, backing, etc.

Figure 3: Schema for NYC Motor Vehicle Collisions - Vehicles

2. NHTSA Safety Ratings of Cars from Department of Transportation

Fieldname	Vehicles	Overall Rating	Frontal Crash	Side Crash	Rollover	Safety Tech
Data Type	int	int	int	int	int	String
Brief description	Vehicles model name	Overall rating of the model	Safety score of the model for frontal crash	Safety score of the model for side crash	Safety score of the model for rollover	Recommended safety technology available for the model

Figure 4: Schema for NHTSA Safety Ratings of Cars

3. NYC Motor Vehicle Collisions - Persons from NYPD

Fieldname	unique_id	person_injury	vehicle_id	pos_in_car	sex
Data Type	int	String	int	String	String
Brief description	Unique record code generated by the system for each person involved in a collision.	Categorical data of whether the person is injured or not.	Foreign key from the NYC Motor Vehicle Collisions - Vehicle data	Description of which seat the injured person is in, such as “passenger” or “driver”.	Gender of the person. Either “F” or “M”.

Figure 5: Schema for NYC Motor Vehicle Collisions - Persons

6 Data Processing and Analytic Stages

Data Scraping from Car Safety Ratings Website (Dataset 2)

The data of each brand safety rating was provided as a chart on the website. To be analyzed with MapReduce, csv files for each brand were needed thus, web scraping was necessary.

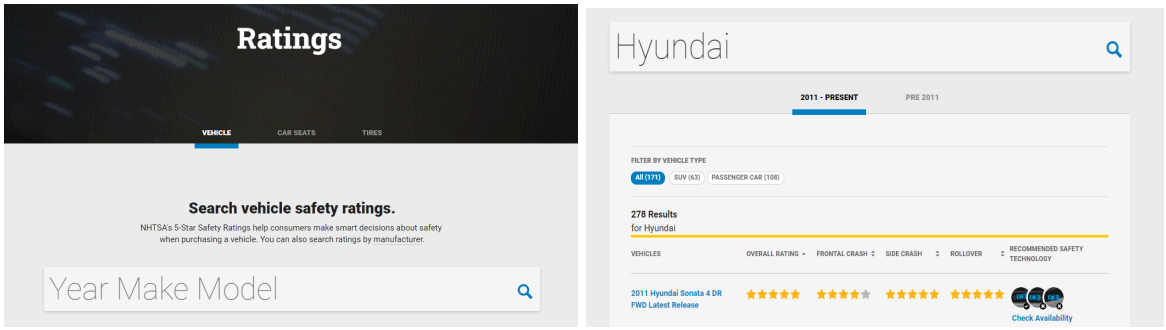


Figure 6 : Web browser of ratings(Department of Transportation https://www.nhtsa.gov/ratings)

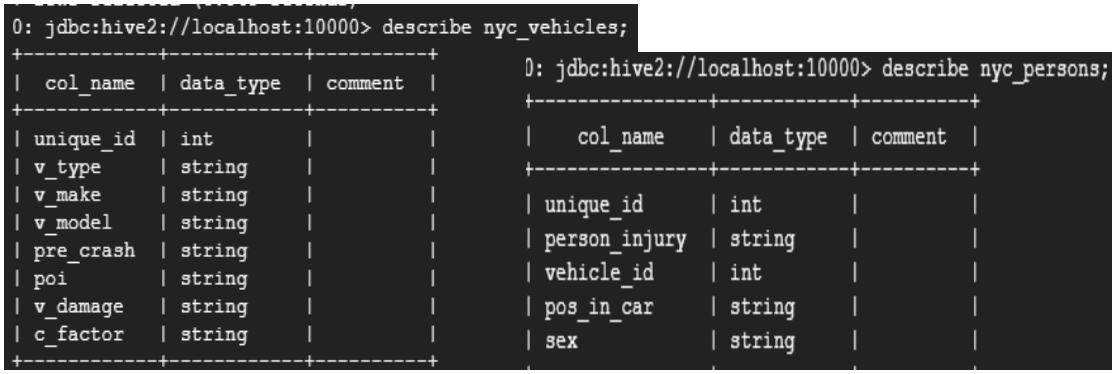
Due to the design of the website, automated interaction with the web browser was needed thus selenium from Python was chosen. With Python, input, search, and scrape of data were automated through the web browser. While scraping, there were null values in the chart which were not wanted due to the data being too old or missing values in numerous columns. Thus, those rows were skipped in the data scraping process through conditional statements.
Then the csv files for each brand were sent to Dataproc to be analyzed to get some key analytical information about each brand through MapReduce.

Data Ingestion/Cleaning for Dataset 1 & 3:

The datasets were downloaded to a local disk and uploaded to HDFS of NYU DataProc via the “put” command. MapReduce was used to transform all strings to lower cases, and only the useful columns described above in part 5 were included to create cleaned csv files. The two cleaned csv files were then imported into HIVE as HIVE tables.

Import cleaned dataset 1 & 3 into HIVE

The cleaned csv files were imported into HIVE using HiveQL statements. Below shows the describe functions of each HIVE table:



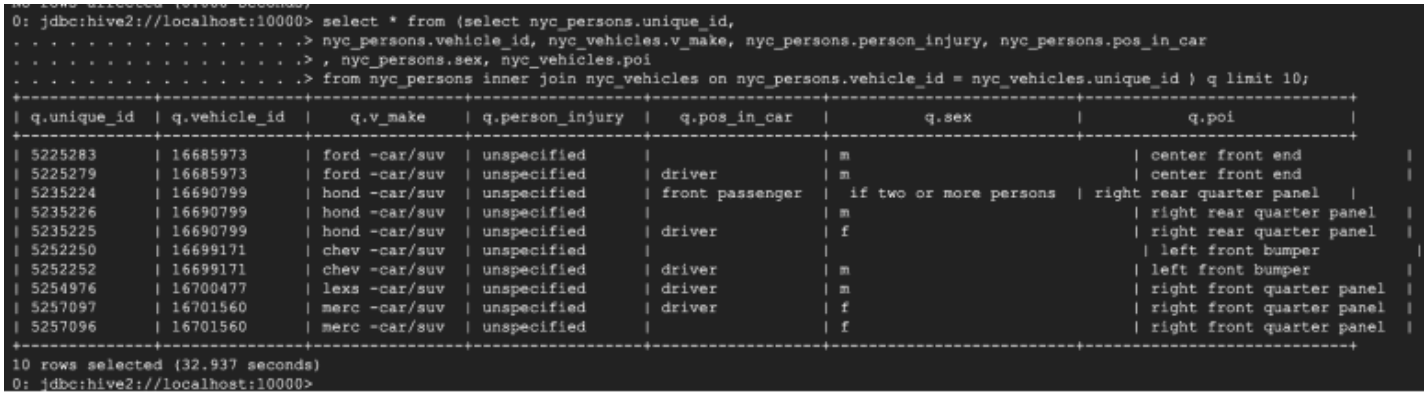
```
0: jdbc:hive2://localhost:10000> describe nyc_vehicles;
+------------+------------+----------+
|  col_name  | data_type  | comment  |
+------------+------------+----------+
| unique_id  | int        |          |
| v_type     | string     |          |
| v_make     | string     |          |
| v_model    | string     |          |
| pre_crash  | string     |          |
| poi        | string     |          |
| v_damage   | string     |          |
| c_factor   | string     |          |
+------------+------------+----------+
```

```
0: jdbc:hive2://localhost:10000> describe nyc_persons;
+----------------+------------+----------+
|    col_name    | data_type  | comment  |
+----------------+------------+----------+
| unique_id      | int        |          |
| person_injury  | string     |          |
| vehicle_id     | int        |          |
| pos_in_car     | string     |          |
| sex            | string     |          |
```

Figure 7 : Table showing cleaned csv of dataset 1 & 3 files into HIVE tables

The two HIVE tables were then joined together using the foreign key in nyc_persons. The “vehicle_id” corresponds to the “unique_id” in the nyc_vehicles HIVE table.

Use Inner Join to join dataset 1 & 3 and obtain insights



```
0: jdbc:hive2://localhost:10000> select * from (select nyc_persons.unique_id,
. . . . . .> nyc_persons.vehicle_id, nyc_vehicles.v_make, nyc_persons.person_injury, nyc_persons.pos_in_car
. . . . . .> , nyc_persons.sex, nyc_vehicles.poi
. . . . . .> from nyc_persons inner join nyc_vehicles on nyc_persons.vehicle_id = nyc_vehicles.unique_id ) q limit 10;
+--------------+---------------+---------------+-----------------+-----------------+-------------------------+-------------------------+
| q.unique_id  | q.vehicle_id  |    q.v_make   | q.person_injury |   q.pos_in_car  |          q.sex          |          q.poi          |
+--------------+---------------+---------------+-----------------+-----------------+-------------------------+-------------------------+
| 5225283      | 16685973      | ford -car/suv | unspecified     |                 | m                       | center front end        |
| 5225279      | 16685973      | ford -car/suv | unspecified     | driver          | m                       | center front end        |
| 5235224      | 16690799      | hond -car/suv | unspecified     | front passenger |  if two or more persons | right rear quarter panel |
| 5235226      | 16690799      | hond -car/suv | unspecified     |                 | m                       | right rear quarter panel |
| 5235225      | 16690799      | hond -car/suv | unspecified     | driver          | f                       | right rear quarter panel |
| 5252250      | 16699171      | chev -car/suv | unspecified     |                 |                         | left front bumper        |
| 5252252      | 16699171      | chev -car/suv | unspecified     | driver          | m                       | left front bumper        |
| 5254976      | 16700477      | lexs -car/suv | unspecified     | driver          | m                       | right front quarter panel |
| 5257097      | 16701560      | merc -car/suv | unspecified     | driver          | f                       | right front quarter panel |
| 5257096      | 16701560      | merc -car/suv | unspecified     |                 | f                       | right front quarter panel |
+--------------+---------------+---------------+-----------------+-----------------+-------------------------+-------------------------+
10 rows selected (32.937 seconds)
0: jdbc:hive2://localhost:10000>
```

Figure 8 : Table showing joined table of dataset 1 & 3

The figure shows what is contained in the joined table. From this joined table, we can use HiveQL to query information and insights about the relationship between point of impact and injury information. We can also query information based on the car brand and obtain injury proportions for each of the car brands.

7 Visualization

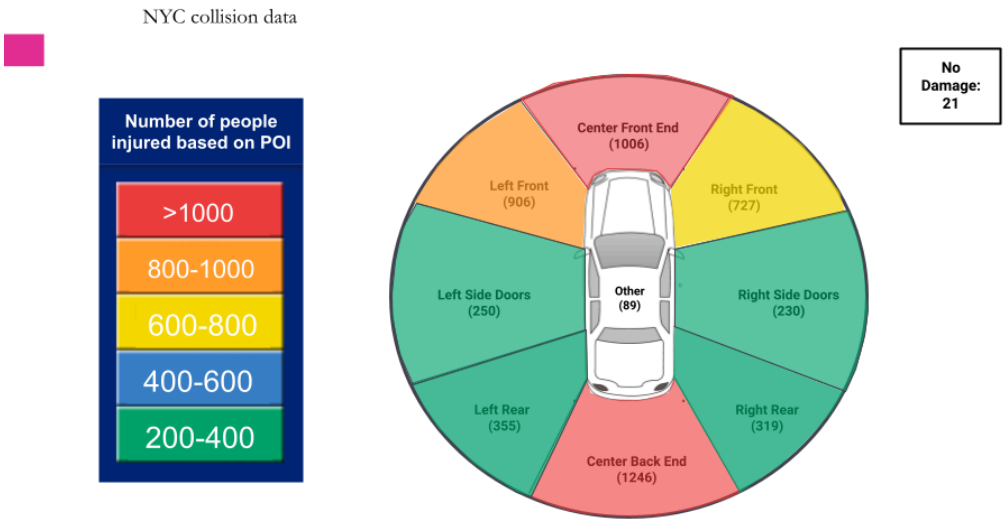


Figure 9 : Diagram showing Relationship between POI and number of people injured.

For initial analysis from the joins table, we can observe that most people suffered injuries from collisions with the point of impact at the Center Front and Center Back ends, whereas fewer people got injured from the side and rear directions. However, we cannot make the statement that frontal or center back ends are leading to more injuries, because it might simply be the case that front and back end collisions are much more common than side collisions.

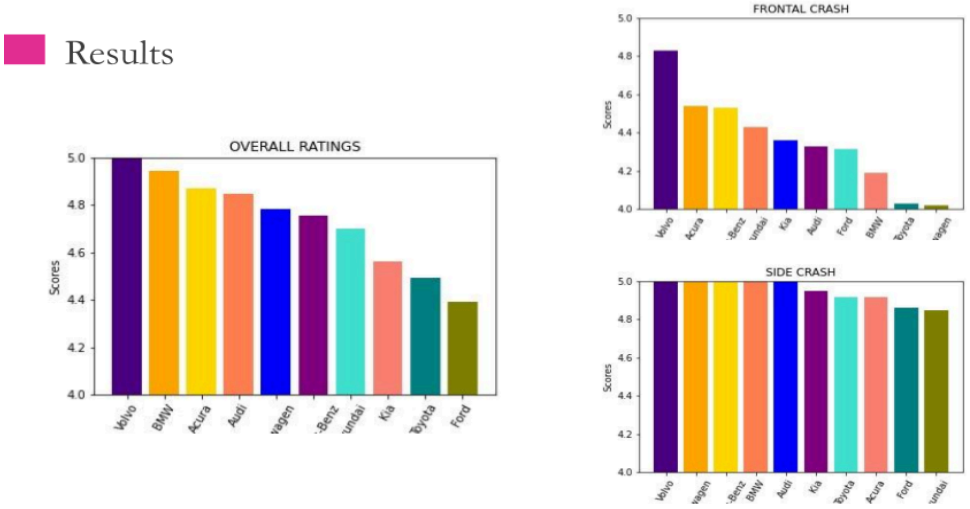


Figure 10 : Bar graphs showing Car Safety Rankings of multiple car brands and POI.

Next, we obtained insight from the visualization of Car Safety Rankings. We can observe that Volvo has the highest ratings in frontal, side, and overall ratings among the popular brands. Especially for frontal crashes, Volvo has a much higher rating compared to all of the other brands.

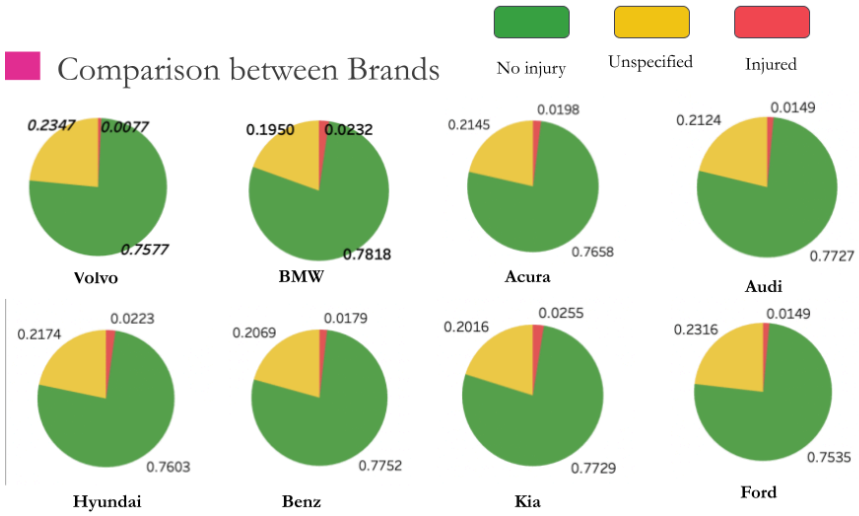


Figure 11 : Pie Charts showing Injury Rates of multiple car brands in NYC collisions data

Finally, this visualization compares the injury rates across the popular brands using data from the NYC collision data. The dataset labels each person as either “no injury”, “unspecified”, or “injured”. We can observe that with an injury rate of 0.77%, Volvo is clearly the winner among these brands. This result is consistent with the previous bar graph showing Volvo as having the highest ratings in the Safety Ratings data.

8 Conclusion

The highlight of findings: From our visualization, we observed that Volvo had the highest ranking for all points of impact out of all these popular car brands in the Car Safety Rankings dataset. Moreover, from the NYC real collision dataset, out of all the car brands and their collision data, Volvo also had the lowest injury rate, which suggests that it is consistent with the safety ratings data and is worth the top ranking.

Implications of study: We have seen previous research papers focusing on case studies of just very few car brands and their safety features. The method we used to generate our results helps to create a bigger picture of comparisons between a car’s safety rating and its injury report in the real world. Since we are using proportions and not absolute numbers, a direct comparison can be done across many different car brands even though the total number of car collisions for various brands is completely different.

Importance of the findings: One implication of our study is to provide a more comprehensive review of a car brand's safety level. If we were to only look at data from the Car Safety Ratings website, those data are only collected from testing fields and not from realistic settings. Since driving on the real road is a much more sophisticated process, people might suspect that the rating system may not truly reflect a car brand's ability to reduce occupant injuries. Thus, by making a direct comparison between real collision data and car safety ratings data, we were able to show that even in the real-world setting, Volvo does an outstanding job of reducing occupant injuries.

New ways of viewing the research problem: Our current method obtains insight using aggregate results for all models of a single brand, but the safety levels of each model within each brand may differ. Thus, one new possible expansion to our research paper is to investigate the injury rate based on specific models of a car brand. (We tried doing this method initially but the NYPD dataset had too little useful information based on specific car models to generate any meaningful insights).

9 References

2018 statistics about NYC and NYC car numbers
https://dmv.ny.gov/statistic/2018reginforce-web.pdf

NYS cars registration
https://data.ny.gov/Transportation/Vehicle-Makes-and-Body-Types-Most-Popular-in-New-Y/3pxy-wy2i

United States Department of Transportation
https://www.nhtsa.gov/ratings

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

Motor Vehicle Collisions - Person
https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu

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

Volvo Drivers' Experiences With Advanced Crash Avoidance and Related Technologies
https://www.tandfonline.com/doi/full/10.1080/15389588.2013.798409

Volvo Cars' Side Impact Protection Systems and their Effectiveness
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3242537/

IIHS Volvo City Safety
https://www.media.volvocars.com/us/en-us/media/pressreleases/173214/iihs-study-volvos-city-safety-reduces-rear-end-crashes-by-41-injuries-by-48