Crash-Safety Ratings and the True Assessment of Injuries by Vehicle

Cody Philips* Robert Garrett† Alan J. Tatro‡ Thomas J. Fisher§

Abstract

Every year the National Highway Traffic Safety Association (NHTSA) and Insurance Institute for Highway Safety (IIHS) release safety ratings for popular makes and models of vehicles produced. We link these safety ratings with the accident data provided in the National Automotive Sampling System (NASS) General Estimates System (GES). We develop a web-based dashboard to graphically explore the relationship among these datasets. We create a metric that measures average injuries per vehicle per accident and map it to vehicle safety ratings. Our web application allows a user to explore different aspect of accidents (e.g., use of alcohol, speeding, etc.) and compare injuries and safety rating performance based on these conditions. Lastly, our dashboard allows a user to see differences between the NHTSA and IIHS ratings systems and how they correspond to the data in NASS GES.

Key Words: Accident Data, Dashboard, Vehicle Safety Ratings.

1. Introduction

The National Highway Traffic Safety Administration (NHTSA) is a federal agency part of the Department of Transportation tasked with saving lives, preventing injuries and reducing vehicle-related crashes; see NHTSA (2016c). Each year, the National Automotive Sampling System (NASS) collects vehicle accident data from police reports nation-wide. In particular, the General Estimates System (GES) within NASS records data from a nationally representative sample of police reported motor vehicle crashes of all types, from minor to fatal; see NHTSA (2016b). Since its inception in 1988, the GES has created a database of nearly 30 years of motor-vehicle accident data including variables on vehicles parameters, personal injuries, and environmental conditions. The GES is similar to the well-known Fatality Analysis Reporting System (FARS) data but includes all levels of vehicle accidents.

Data in the GES is sampled from police accident reports that involve at least one motor vehicle traveling on a traffic way, and the result of the accident must be property damage, injury, or death. The accident reports are sampled from 60 areas that reflect the geography, roadway mileage, population, and traffic density of the United States. Approximately 50,000 police accident reports are collected each year and included in GES. The raw GES data are available: `ftp://ftp.nhtsa.dot.gov/GES/`.

Vehicle safety is a primary criterion for many when purchasing a new car. Since 1978, the NHTSA has implemented the so-called "5-Star Safety Ratings Program" for motor-vehicles, wherein vehicles undergo a series of crash-safety tests and are assigned a qualitative measure of safety (1

*Corresponding author Cody Philips is a Senior Undergraduate Student in Mathematics & Statistics with an Analytics Co-Major at Miami University, contact information: Department of Statistics, Miami University, 311 Upham Hall, Oxford, OH 45056 USA, Tel: 1-513-529-7828, Fax: 1-513-529-0989, Email: philipcr@miamioh.edu

†Robert Garrett is a Sophomore Undergraduate student in Mathematics & Statistics with an Analytics Co-Major at Miami University, Email: garretrc@miamioh.edu

‡A.J. Tatro is a Senior Undergraduate student in Mathematics & Statistics with an Analytics Co-Major at Miami University, Email: tatroaj@miamioh.edu

§Faculty Advisor Thomas J. Fisher is an Assistant Professor in the Department of Statistics at Miami University, Email: fishert4@miamioh.edu

star is the lowest level with 5 being the safest); see NHTSA (2016a). Meanwhile, the independent, nonprofit and educational organization, the Insurance Institute for Highway Safety (IIHS) performs tests on vehicles *crashworthiness* which is defined as how well a vehicle protects its occupants in a crash; see IIHS (2016). Although the testing mechanisms and focus of the NHTSA and IIHS ratings are different, each provides a measure of *safety* to a (potential) consumer of a particular vehicle.

In this project, we link the "5-Star" vehicle safety ratings issued by NHTSA and the safety ratings issued by the IIHS to the vehicle records within the GES from 2002 through 2014. By linking these different sources of data, we can inspect the performance of different vehicles in motor-vehicle accidents and cross-check it against its safety ratings. The outline of this article is as follows: Section 2 outlines the arduous task of collecting and linking the three sources of data. Section 3 describes the implementation of visualization tools via a *shiny* dashboard web app. Lastly, Section 4 provide some conclusions and discussion related to the data.

2. Data Process

2.1 NASS GES Data

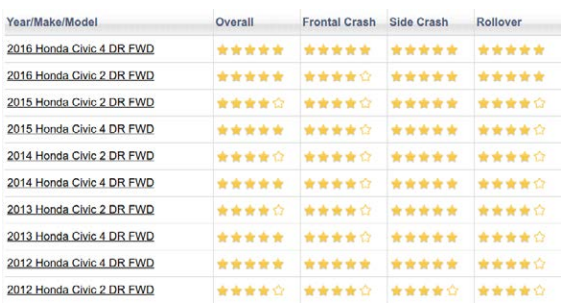
In this project we used the NASS GES data from 2002 through 2014 (we typically refer to it as the *NASS data*). The NASS data was processed in SAS 9.4 using the supplied source code and SAS formatted data files (.sas7bdat files for each year of data). We chose data from 2002 through 2014 in order to maintain consistency for our analysis, as years before 2002 had great discrepancies in the variables recorded and the documentation involved from the 2002 through 2014 datasets. The end goal of the project was to implement a dynamic user interface in R linking the vehicle information with injury reports (and accident parameters) to the vehicle safety ratings. After we stripped the SAS specific formatting of the variable, we exported the *accident* and *vehicle* SAS datasets into csv files to import into R (R Core Team, 2015) using the `data.table` package (Dowle et al., 2015).

Within the GES data, specific vehicles are indicated by year and codes for the make and model (e.g., the 2012 Honda Civic is marked with make code 37, model code 31 and year 2012). The model codes can be found in the 1988–2007 GES Analytical User Manual. The vehicle model code tables were extracted using Adobe Acrobat (need to check with Vickie) into a form suitable for Microsoft Excel. Once in Excel, some further processing created a table matching the vehicle codes with their named Make and Model using fuzzy matching.

2.2 Processing Safety Data

Both the NHTSA and IIHS provide web interfaces that allow consumers to search vehicle safety ratings. Both websites provide graphical outputs reporting vehicle safety. Figure 1 provides example output for the Honda Civic over a select number of recent years. Several computational challenges can be seen from the example. Both the NHTSA and IIHS websites report the safety ratings in graphical form. Although both are measuring *safety*, the specific ratings reported are different. NHTSA provides an "overall" category but the IIHS ratings are for specific categories. Further, the scaling of their ratings are different: NHTSA uses a 5-star approach with 1 corresponding to the lowest rating and 5 being the best; IIHS rates by labeling Good (G), Acceptable (A), Marginal (M) or Poor (P).

Both the NHTSA and IIHS were contacted in an effort to obtain their full sets of safety ratings data. For both agencies, they encouraged we obtain the safety ratings through the designated web applications. We used the Eclipse Java IDE and the UserClient class of the Jaunt API to read the webpage directory in HTML format. We created a new UserClient, which looped through each page



Year/Make/Model	Overall	Frontal Crash	Side Crash	Rollover
2016 Honda Civic 4 DR FWD	★★★★★	★★★★★	★★★★★	★★★★★
2016 Honda Civic 2 DR FWD	★★★★★	★★★★☆	★★★★★	★★★★★
2015 Honda Civic 2 DR FWD	★★★★☆	★★★★☆	★★★★★	★★★★☆
2015 Honda Civic 4 DR FWD	★★★★★	★★★★☆	★★★★★	★★★★☆
2014 Honda Civic 2 DR FWD	★★★★☆	★★★★☆	★★★★★	★★★★☆
2014 Honda Civic 4 DR FWD	★★★★★	★★★★☆	★★★★★	★★★★☆
2013 Honda Civic 2 DR FWD	★★★★☆	★★★★☆	★★★★★	★★★★☆
2013 Honda Civic 4 DR FWD	★★★★★	★★★★☆	★★★★★	★★★★☆
2012 Honda Civic 4 DR FWD	★★★★★	★★★★★	★★★★★	★★★★☆
2012 Honda Civic 2 DR FWD	★★★★☆	★★★★☆	★★★★☆	★★★★☆

(a) NHTSA 5-star safety rating

Model year	Front overlap		Side	Roof strength	Head restraints & seats	Front crash prevention
	Small	Moderate				
2016	G	G	G	G	G	superior
2015	G	G	G	G	G	basic
2014	G	G	G	G	G	basic
2013	G	G	G	G	G	basic
2012		G	G	G	G	
2011		G	G	G	G	
2010		G	G	G	G	

(b) IIHS Crashworthiness

Figure 1: Safety ratings for select years of the Honda Civic as reported by NHTSA 5-star safety ratings and IIHS crashworthiness websites.

```
<th class="fill" scope="row"><a href="http://www.safercar.gov/Vehicle+Shoppers/5-Star+Safety+Ratings/1990-2010+Vehicles/Vehicle-
Detail?vehicleId=5834" id="5834">2010 Honda Civic 2-DR. w/SAB</a></th>
<td class="stars b_right"><img src="http://www.safercar.gov/staticfiles/scassets/images/stars_5.png" alt="5 star rating" class="stars"></td>
<td class=" topalign"> </td>
<td class="stars b_right"><img src="http://www.safercar.gov/staticfiles/scassets/images/stars_5.png" class="stars" alt="5 star rating"></td>
<td class=" topalign"> </td>
<td class="stars b_right"><img src="http://www.safercar.gov/staticfiles/scassets/images/stars_4.png" class="stars" alt="4 star rating"></td>
<td class=" topalign"> </td>
<td class="stars b_right"><img src="http://www.safercar.gov/staticfiles/scassets/images/stars_5.png" class="stars" alt="5 star rating"></td>
<td class=" topalign"> </td>
<td class="stars b_right"><img src="http://www.safercar.gov/staticfiles/scassets/images/stars_4.png" class="stars" alt="4 star rating"></td>
<td class=" topalign"> </td>
<td class="stars b_right"><img src="http://www.safercar.gov/staticfiles/scassets/images/stars_not_tested.png" class="stars" alt=" star
rating"></td>
<td class=" topalign"> </td>

2010 Honda Civic NHTSA Data (Text)
```

Figure 2: Example of raw text file after web scraping

of the user view of the NHTSA and IIHS websites and paste all the content into a text document. An example of this output can be seen in Figure 2. From the resulting file, safety rating information could be extracted. Another Java application was written to clean the text document by selectively gathering vehicle names and the corresponding image file names (JPG files corresponding to 4-star, 5-star, etc.) that revealed the rating of the car. The raw data was then converted into a csv file for use in R and example can be seen in Figure 3.

Once the entire database of safety ratings (NHTSA and IIHS) were available, we needed a way to link them to the GES data. The GES data refers to vehicles by their make and model codes and year, so for popular vehicles we were able to write a script appending the make and model codes from the GES data to the two sets of safety ratings. For the remainder of vehicles, manual brute force entry of linking was achieved through fuzzy matching of vehicle makes and models.

As aforementioned, the NHTSA 5-star safety ratings include an "overall" category but the IIHS ratings are for specific vehicle safety test. Further, in some cases the NHTSA provides multiple ratings for the same make and model whereas the GES data does not distinguish between vehicle class options; for example, in Figure 1a we see different ratings for the two door and four door 2013 Honda Civic while the GES data may just report a 2013 Honda Civic. To provide one overall safety

```
2010 Honda Civic 2-DR.,5star,5star,4star,5star,4star,no rating
```

Figure 3: Example of csv format after web scraping and processing

rating for each vehicle make and model we used the floor function of the median overall 5-star rating (in our example, the 2013 Honda Civic is labeled with 4 stars). To compute an "overall" rating for the IIHS ratings, we used the modal response of all available safety ratings. In our example in Figure 1b, all Honda Civics receive a *Good* overall rating.

2.3 Linking the Data

Once all the preliminary data processing was complete, the NASS and Crash-safety ratings were merged in R (R Core Team, 2015). Many different functions and packages were used due to the diverse nature and overall structure of the datasets involved, but much of the processing relied on the `dplyr` package (Wickham and Francois, 2015).

The initial reading in of the data was a bit daunting, as there are 26 files to read in each ranging from 7MB to 15 MB in size. To remedy the need for individual file input, we use the `list.files` function to grab the names of all the files in the folder and load them into R.

The next step was to merge the accident and vehicle datasets together in each year in an effective and efficient manner. We performed a *left outer join* with the vehicle dataset as the base and the accident dataset being merged. This *join* type was used because our analysis concentrates on vehicle performance. An outer join allowed us to retain all vehicles even if a particular vehicle did not record an accident in a given year. The `merge` function in R handles these types of merges incredibly well.

After merging each year to get vehicle-level data on accidents, we select the variables of interest Some processing using the `data.table` package (Dowle et al., 2015) allows the changing of variable names. For example, in the 2008 data, the NASS dataset changed the following variables: 'SPEED' into 'TRAV_SP,' 'VEH_SEV' into 'DEFORMED,' and in 2011 the NASS dataset changed the variable 'MOD_YEAR' into 'MODEL_YR.' These variables were changed as there are slight changes in the recorded levels for each of these variables. The processing of these changes is outlined later in the paper.

The most important step in our processing is creating the primary key to link the NASS data to the safety ratings data from the web scraping and excel data cleaning process. We created a concatenated form of the make, model, and model year variables in both datasets. We used the `paste` function to concatenate with a space in between each variable to be a part of the primary key. Ultimately, this allows each vehicle in the NASS dataset to be properly linked with its corresponding safety rating.

Once all rows of the NASS data and vehicle safety ratings were processed, the data for each year of study were stacked using the `rbind` function, we merged the NASS data with the vehicle safety ratings dataset using a *left outer join* to ensure we are still retaining vehicles without any accidents in the NASS data but with either IIHS or NHTSA ratings. Finally, we get to the more intricate parts of the data cleaning process. Using the `dplyr` package, and its various SQL-like commands, the process for creating groupings of vehicles and injury summaries is accomplished through functions such as the `select`, `group_by` and `summarise`. These techniques allows us to compare average major injury (the MAX_VSEV variable), the average number of injuries per accident using (the NUM_INJV variable).

We then create two `data.frames`. The former data (grouped by the make, model and year) is grouped by IIHS and NHTSA ratings, respectively to get the average injury sustained by crash-safety rating, by vehicle, by accident. This format of the data was more viable for our implementation of the data visualization discussed later.

2.3.1 Changes within the NASS data

As aforementioned, several of the variables in the NASS data changed format and needed to be updated to reflect the more rounded view of the entire dataset. A couple of key assumptions are to be noted: MAX_VSEV had values of 6, which used to be coded as "died prior [to crash]," so this needed to be coded back to a value of 4, in order to represent a more general level of "Fatal Injury." We removed all unknown or missing values in every variable that had such possible values. This was done because with data on a per vehicle level, it is unclear how to deal with unknown values. Lastly we switched the values of 4 and 5, since 4 was originally "Fatal Injury" and 5 was originally "Injured Severity Unknown." This decision was made to ensure a more ordinal understanding of the variable. An injury with severity unknown is less intense than a fatal injury, and thus this seemed to be an appropriate decision.

Since the speeding variables changed over time in the NASS data, we created our own speeding metric. From the accident vehicle dataset we have the speed at which the vehicle was going, as well as the speed limit in the area. Thus we can calculate a new variable "isSpeed," as an indicator variable of whether the vehicle was speeding (that is, a 1 when the travel speed is greater than the speed limit, and a 0 otherwise).

3. Visual Analysis

Our analysis of the NASS GES vehicle data is largely accomplished through a `shinydashboard` app (Chang et al., 2016; Chang, 2016) that utilizes `ggplot2` (Wickham, 2009) functionality. This provides the user with some interactivity to explore the database and safety ratings. The application can be accessed through the link: `http://dataviz.miamioh.edu/nhtsaGesAnalysis/`. Below we describe specific parts of the shiny dashboard and discuss some of the findings.

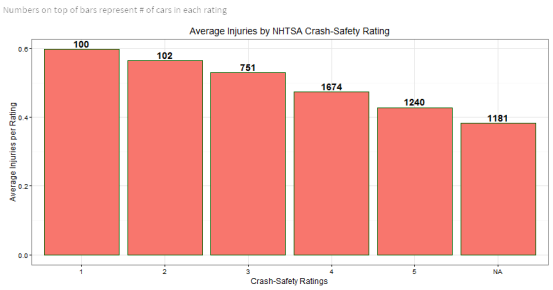
3.1 Overview Tab

The first tab in the app is the Overview tab, which displays static plots of an overview of the data. This provides a general understanding of the research problem of study. Figure 4 displays the average injury by crash-safety rating for both the NHTSA and IIHS ratings. We chose to display the average injuries with a bar graph labeled with the size on the top of the bar because some specific vehicles (when using the *Make* or *Manual Entry* tab) have very few entries (certain vehicles may only have a single entry in the entire NASS dataset) and other plots breakdown in these extreme cases. Although not visually optimal, the simple bar graph approach works nicely in all cases.

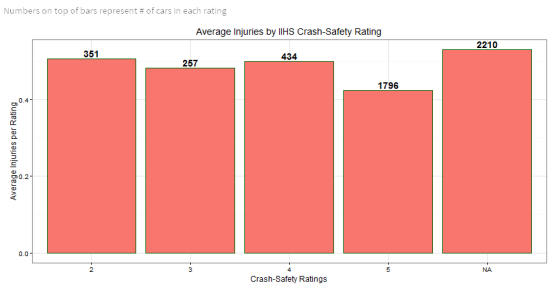
We can see in Figure 4 as the vehicles are rated safer by the NHTSA the average injuries decreased. This is what we would expect from a consumer standpoint. However, IIHS ratings do not align as much with our preconceived notions of crash-safety ratings; perhaps due to our aggregated "overall" IIHS rating. There is not much difference between 2, 3, and 4 crash-safety ratings (2 - Poor, 3 - Marginal, 4 - Acceptable, 5 - Good) in terms of average injuries sustained, while those receiving a 5 is considerably lower.

3.2 Make Overview

The next section is the Make Overview tab on the shiny dashboard. The purpose of this option is allow a user to generate similar plots to the overview plot, but conditional on a particular make. This provides consumers an understanding of the reality of vehicle accident data analyzed directly against how safe a car is said to be. Figure 5 below provides a view of the Ford cars by NHTSA crash safety rating and Chevrolet cars by IIHS crash-safe

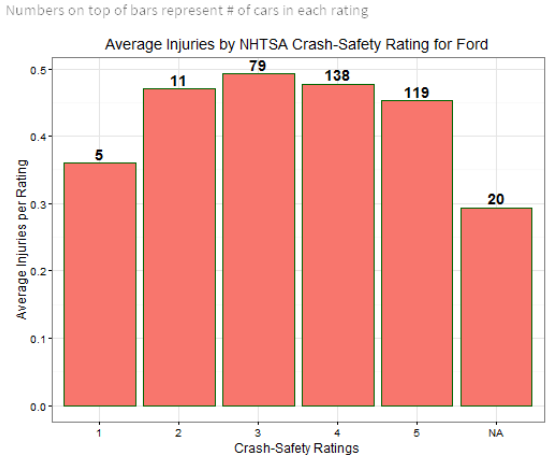


(a) NHTSA 5-star safety rating

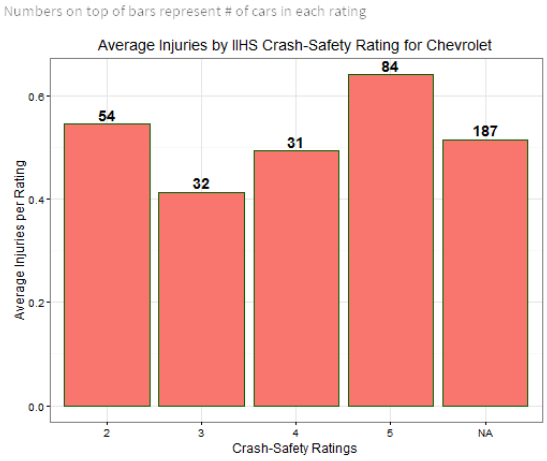


(b) IIHS Crashworthiness

Figure 4: Average number of injuries per accident as a function of vehicle safety ratings (NHTSA on left, IIHS on right (2 - Poor, 3 - Moderate, 4 - Acceptable, 5 - Good)). The value at the top of each bar provides the sample size of records for a given average.



(a) Ford NHTSA 5-star safety rating



(b) Chevrolet IIHS Crashworthiness

Figure 5: Average number of injuries per accident as a function of vehicle safety ratings (NHTSA on left, IIHS on right (2 - Poor, 3 - Moderate, 4 - Acceptable, 5 - Good)) by vehicle *Make*. The value at the top of each bar provides the sample size of records for a given average.

As we can see from Figure 5, the NHTSA crash safety-ratings for Ford, for the most part, follow along the trend we see in the overview plots. Moreover, it is interesting to note that while the average injury does decrease from 3 to 5 star NHTSA Ratings for Ford, they do not decrease much. If safety is the highest priority in a car, but the 5 star Ford car is more expensive (as 5 star rated cars usually are), it might make financial sense to purchase the 3 or 4 star car. According to the data in the NASS system, you are not increasing average injury by much when moving from a 5 star Ford car to a 3-star Ford car. The IIHS crash safety-rating plot may provide an example of the limitations of the NASS (it is a sample) dataset as our result is not necessarily intuitive. For Chevrolet vehicles, the cars that are rated as Good by the IIHS rating system had the highest average injury rate. This can likely be explained as an issue with variability within the safety ratings (there were 85 vehicles in the Good rating compared to just 32 in the Moderate rating). Lastly, we remind the reader these are just two examples of the many possible vehicle Makes that can be analyzed in the system.

Figure 6: User options within the Manual Entry Tab

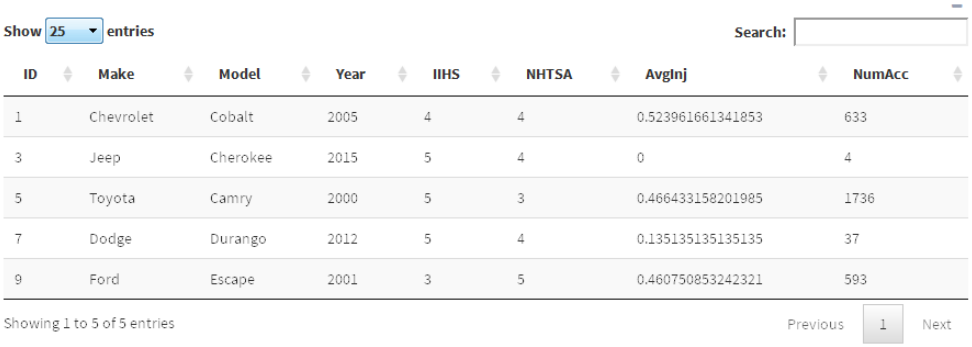


Figure 7: Example output of comparing multiple vehicles in Manual Entry Tab

3.3 Manual Tab

The next tab in the shiny dashboard is the manual entry tab. This provides sample statistics for a specific vehicle make, model and year as selected by the user. The datatable gives us the NHTSA and IIHS crash-safety ratings, as well as the average injury amount sustained and the number of accidents the NASS data has available for this specific vehicle. Figure 6 provides a view of the manual input options one can select.

By choosing several vehicles we can explore the performance of individual cars side-by-side. This allows a user to make decisions specifically between multiple cars in buying preference, and/or safety analysis. Figure 7 provides an example of a set of potential selections of multiple vehicles and its datatable output.

In Figure 7 we see that some cars have considerable amount of accidents recorded in the NASS data; for example, the 2000 Toyota Camry had 1736 accidents in the database with an average injury level of 0.466. However, the 2015 Jeep Cherokee only has 4 accidents recorded and none of them had an injury sustained. This example further demonstrates the limitations of the NASS database when comparing individual vehicles. However, it also provides an example of the power of data and our shiny dashboard. The 2001 Ford Escape and 2005 Chevrolet Cobalt have a comparable number of accidents (593 and 633, respectively) but have different average injury metrics and safety ratings. In fact, the NHTSA safety ratings appear to more closely match our injury metric (higher safety

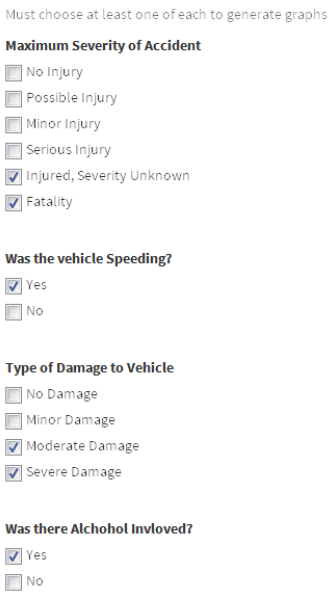


Figure 8: Options available in Accident Condition tab

rating corresponding to less injuries).

3.4 Accident Condition tab

The final tab allows the user to block on certain conditions that were part of the accident. Each box that is checked creates a different reactive data-frame and plots a new graph that is reflective of the conditions chosen. The manual check box options are shown in Figure 8. In that example, we are conditioning on the severity of the accident (in terms of sustained injury), whether speeding was involved, vehicle damaged and the use of alcohol.

Figure 9 provides the average injuries sustained as a function of the NHTSA and IIHS safety ratings. The obvious finding of this example is when a vehicle was speeding, and there was alcohol involved, as well there was a fatality or an unknown injury and the maximum damage to each was moderate or severe, then there is a fairly high average injury rating regardless of vehicle safety rating. The interesting fact to note is that the average injury does not change much in these severe conditions. If a vehicle is speeding and alcohol is involved, the crash-safety rating of a vehicle does not have much influence on average injuries.

4. Discussion

Overall, the NASS GES dataset provides some understanding of the reality of vehicle safety. When augmented with government (NHTSA) and the insurance (IIHS) agencies, we can gain some interesting insights about vehicle safety. The results of this project suggest that the NHTSA crash-safety rating is more credible than the IIHS rating. If someone wants to truly understand the safety of a particular vehicle, the overall NHTSA crash-safety rating will provide a good approximation for such an inquiry.

Additionally, our data processing and shiny dashboard allow users to explore other aspects of vehicle safety. Many people trust particular brands or makes due to brand loyalty or pre-conceived

(a) Ford NHTSA 5-star safety rating

(b) Chevrolet IIHS Crashworthiness

Figure 9: Average number of injuries per accident as a function of vehicle safety ratings (NHTSA on left, IIHS on right (2 - Poor, 3 - Moderate, 4 - Acceptable, 5 - Good)) under certain accident conditions: alcohol involved, speeding and moderate to severe accident. The value at the top of each bar provides the sample size of records for a given average.

notions. Our manual entry tool allows the user to select an individual make and analyze its safety performance against other vehicles. Our findings are not conclusive however, as further work could involve creating a more rounded "overall" metric for IIHS ratings or allowing a user to pivot on specify safety metrics (e.g., Frontal Crash ratings). Further our work can be expanded to create a predictive model to assess how likely a given vehicle will perform (in terms of safety) when faceting on various vehicle attributes and accident conditions.

Acknowledgments

The research team would like to thank the ASA Government Section on Statistics for sponsoring the Data Science Challenge. The lead author is appreciative of a travel grant from the Office of Research for Undergraduates at Miami University. Lastly, the research team is thankful to Vickie Sandlin, Program Associate for the Department of Statistics at Miami University, for extracting the NASS vehicle codes from the GES Analytical Users Guide from PDF format to Word format.

References

Chang, W. (2016), *shinydashboard: Create Dashboards with 'Shiny'*, r package version 0.5.3.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2016), *shiny: Web Application Framework for R*, r package version 0.13.2.

Dowle, M., Srinivasan, A., Short, T., with contributions from R Saporta, S. L., and Antonyan, E. (2015), *data.table: Extension of Data.frame*, r package version 1.9.6.

IIHS (2016), "Safety Vehicle Safety Ratings," `http://www.iihs.org/iihs/ratings`, accessed: 2016-09-14.

NHTSA (2016a), "5-Star Safety Ratings," `http://www.safercar.gov/Safety+Ratings`, accessed: 2016-09-14.

— (2016b), "NASS General Estimates Systems," `http://www.nhtsa.gov/Data/National+Automotive+Sampling+System+(NASS)/NASS+General+Estimates+System`, accessed: 2016-09-14.

— (2016c), "National Highway Traffic Safety Administration," http://www.nhtsa.gov/, accessed: 2016-09-14.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

Wickham, H. and Francois, R. (2015), *dplyr: A Grammar of Data Manipulation*, r package version 0.4.3.