Computational Statistics (2021) 36:1553–1560
https://doi.org/10.1007/s00180-021-01076-5

EDITORIAL

# The 2016 *Data Challenge* of the American Statistical Association

**Roya Amjadi**[1] · **Wendy Martinez**[2]

Received: 12 January 2021 / Accepted: 16 January 2021 / Published online: 24 February 2021


**Abstract**
The *Sections on Statistical Graphics and Statistical Computing* of the *American Statistical Association* have a long history of issuing *Data Challenges* with the first one starting in 1982/1983. The challenge is now an annual event where most of them use data collected and disseminated by the U.S. government. The data set for the *2016 Data Challenge* was the *Department of Transportation's General Estimates System*. The GES is collected by the *National Highway Transportation Safety Administration* and is a representative sample of police-reported motor vehicle crashes. This editorial introduces the five papers submitted by contestants in the data challenge.

## 1 Introduction

The *Sections on Statistical Graphics and Statistical Computing* of the *American Statistical Association* (ASA—see https://www.amstat.org/) have a long history of issuing *Data Challenges* with the first one starting in 1982/1983 (http://stat-computing.org/dataexpo/). These were originally called *The Data Exposition*, which was soon shortened to the *Data Expo*. Recent contests were called the *Data Challenge* and were organized by the *Government Statistics Section* of the ASA, where government data sets were used for analysis. The now annual *Data Challenge Expo* is jointly sponsored by three sections of the *American Statistical Association*: *Statistical Computing, Statistical Graphics*, and *Government Statistics*. The first joint *Data Challenge* took place in 2016 with the contestants presenting their results at

✉ Wendy Martinez
martinez.wendy@bls.gov

Roya Amjadi
roya.amjadi@fhwa.dot.gov

[1] Federal Highway Administration, Turner-Fairbank Highway Research Center, Office of Safety Research and Development, 6300 Georgetown Pike, McLean, VA 22101, USA

[2] Mathematical Statistics Research Center, Office of Survey Methods Research, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington, DC 20212, USA

the *Joint Statistical Meetings* (JSM) in Chicago, Illinois (July 30 through August 4, 2016).

These data expositions and challenges offer an important service to our communities. They provide data and research questions for students and professionals that will result in papers and presentations. They give faculty real-world problems for their students to tackle, and many entries through the years are based on student projects. They can even result in contestants making important connections with professional contacts who provided the data.

In general, the *Data Challenge Expo* is always open to anyone who is interested in participating. This includes college students and professionals from the private or public sector. The annual contest challenges participants to analyze a data set using statistical and visualization tools and methods. The data sets used in the challenges are often obtained from public US Federal government data sources, as was the case in the 2016 GSS *Data Challenge*. The data set used in the 2016 GSS *Data Challenge* was provided by the *US Department of Transportation*, as described in the next section. There are two award categories in the contest—one for professionals and one for students.

The guest editors of this special issue are grateful for the support of Springer and the editors of *Computational Statistics* for enabling us to publish refereed articles from *Data Challenge* contestants. This special issue for the 2016 GSS *Data Challenge* is the fourth in a hopefully continuing series of special issues in *Computational Statistics* focused on the *ASA Data Challenge Expo*. The first issue focused on the 2006 *Data Expo* where the data set contained geographic and atmospheric measures on a coarse regular grid covering Central America (Murrell 2010). The next special issue covered the 2011 *Data Expo* and pertained to a timely topic—the *Deepwater Horizon* oil spill. The data set consisted of data resulting from monitoring of water temperature and salinity, water chemistry, and relevant wildlife counts (Cook 2014). This was followed by the 2013 *Data Expo* with data from the *Knight Foundation* (www.knightfoundation.org) describing the emotional attachment of residents to their communities (Hofmann et al. 2019).

## 2 The challenge

The *Federal Highway Administration* (FHWA), US states' *Department of Transportations*, and other transportation agencies use data analysis to keep national highways safe and operational. For highway safety and operation applications statistical methodologies are regularly used for analyzing crash data. These methodologies are limited in number, and have limitations in capability and applicability for highway applications. FHWA has identified the need for working closer with statistical communities to advance highway safety and operation research, and practice. There are many reliable national data resources that FHWA can share with statisticians, and benefit from their contributions to highway transportation science. The 2016 GSS *Data Challenge* offered the FHWA opportunities for: gauging statistician's interest for using crash data; and identifying new applicable statistical methodologies that transportation agencies could add to their toolbox.

The data set for the *2016 GSS Data Challenge* was the *Department of Transportation's General Estimates System* (GES). The GES is collected by the *National Highway Transportation Safety Administration* (NHTSA). The GES is a representative sample of police-reported motor vehicle crashes of all types. Crashes recorded in the GES database must involve at least one motor vehicle traveling on a traffic way, have a corresponding police accident report (PAR), and result in property damage, injury, or death. Because GES data are obtained from a probability sample of police-reported crashes, each crash included in the database is associated with a sampling weight, which can be used to obtain population estimates. The data encompass the years from 1988 to 2013. A link to the website with the data and other useful information (data dictionaries, manuals, etc.) is given here: https://www.nhtsa.gov/node/97996/256.

We are grateful for all those who entered the contest. We had seven entries in the Student category and six in the Professional group. The list of 2016 winners are given here, where authors are given as they appear in the JSM 2016 program.

### 2.1 Student category

First place: Ryan Jarrett and Lucy D'Agostino McGowan, "Assessing the Association between Accident Injury Severity and NCAP Car Safety Ratings".
Second place: Aditi Pradeep Sharma, Michael Wierzbicki, and Gaurav Sharma, "Predictive Modeling of Severity of Injuries in Motor Vehicle Crashes".
Third place: Tony Ng, Lynne Stokes, Yifan Zhong, Robert Farrow, Clayton Moore, Gunes Alkan, Haichen Liu, Ziyuan Xu, Yihan Xu, and Yuzhi Yan "Predicting the Potential Economic Cost of a Car Accident under Different Circumstances."

### 2.2 Professional category

Chris Eshleman, Jonathan Auerbach, and Rob Trangucci "Accidents, Injuries, and Driving Speeds: A Causal Investigation".
Abstracts for all entries presented at the JSM can be found here https://ww2.amstat.org/meetings/jsm/2016/onlineprogram/ActivityDetails.cfm?SessionID=213078 and https://ww2.amstat.org/meetings/jsm/2016/onlineprogram/MainSearchResults.cfm.

## 3 Summary of papers in this special issue

We gave all contestants an opportunity to eventually submit a paper to this special issue of *Computational Statistics*. All winners of the *2016 GSS Data Challenge* were invited to contribute a paper, but only two of the winners did so. One is the paper by Gunes et al. (2021) (third place in the student category), and the other is the paper by Jonathan Auerbach et al. (2021) (professional category). Those who did not win were asked to submit a paper to the JSM 2016 Proceedings,

which were then reviewed by the Guest Editors. Three of the papers were deemed as being potentially suitable for publication in this journal and were invited to submit a paper. After going through this process, we ended up with five papers, which are summarized below.

A team of students from Southern Methodist University provided a winning entry in the student category. They described an interesting way to use these data—as a class project. Professors H. K. Tony Ng and Lynne Stokes (Gunes et al. 2021). Ng et al. (2021) led this group of students to victory, and it was a pleasure to see such enthusiastic participants present their work at the *Joint Statistical Meetings*. Their analysis explored factors that might contribute to the level of injury suffered by passengers in a car accident. Using data from the *General Estimation System* and other sources (e.g., car safety ratings and number of fatal crashes per state), they explored factors such as speed, road surface conditions, age, alcohol involvement, and more. The students developed an interactive system called the *Accident Price Explorer* (ACE) and made it available to the public (http://gessmu.azurewebsites.net).

The winning team in the professional category had team members Jonathan Auerbach, Christopher Eshleman, and Rob Trangucci (Auerbach et al. 2021). Their paper explored the issue of selection bias in estimating the effects of traffic safety policy with the *Vision Zero* strategy implemented in *New York City* as a motivating example. *Vision Zero* refers to the strategies established by twelve cities in the United States that will induce drivers to make safer decisions. From their website (https://visionzeronetwork.org/): "The Vision Zero Network is a collaborative campaign helping communities reach their goals of Vision Zero—eliminating all traffic fatalities and severe injuries—while increasing safe, healthy, equitable mobility for all." The authors' particular focus is to explore whether or not selection bias tends to overestimate the benefits of the policy, with the strategy employed in *New York City* as an example. As with the previous paper, these authors also combine data from the *National Automotive Sampling System* (NASS) *General Estimation System* with other databases to improve their Bayesian hierarchical models. Based on the models, the authors conclude that the effects of the policy were overestimated.

The paper by Patrick Coyle, Chen Chen, and Nooreen Dabbish (Coyle et al. 2016, 2021) on drowsy driving addresses an important problem in driving safety. There were approximately 600 incidents of drowsy driving reported by police in 2013. The authors analyze factors such as age, gender, time of day, and the day of the week to explore how drowsy driving might change among sub-populations based on these characteristics. An R package and a Shiny app were developed to allow others to perform similar explorations. The Shiny app can be accessed at https://patrickcoyle.shinyapps.io/GES_plotter/. The appendix in the article shows how to install the R package from the author's github site.

It is important for policy makers to have the right tools to first make policy and then to evaluate the results. The paper by Dooti Roy, Ved Deshpande, and M. Henry Linder (Roy et al. 2016, 2021) focused on accidents involving bus crashes, and they developed a taxonomy of such crashes by applying a two-stage cluster (or unsupervised learning) approach. The first stage used self-organizing maps,

which had the effect of reducing the dimensionality. A neural gas algorithm was used in the second stage. The authors found four distinct clusters that remained stable over time, which is an indication that the groups found were genuine.

The last paper in this special issue was written by students Cody Philips, Robert Garrett, Alan Tatro and their advisor Thomas Fisher (Philips et al. 2016, 2021). They took an interesting approach to the *2016 GSS Data Challenge* by linking safety ratings issued by the *National Highway Traffic Safety Association* and the *Insurance Institute for Highway Safety* with the data on crashes indicated in the GES. One of the main goals of their analysis is to determine how the safety ratings serve as predictors of whether or not someone in a vehicle will be injured in a car crash. The team developed a dashboard to enable the exploration between the data sets via graphs and visualizations. The application also provides the means to explore other factors of the car crashes such as alcohol use, speed, and road type.

## 4 Other submissions to the *2016 GSS data challenge*

Not every contestant in the *Data Challenge* submitted a paper to this special issue. However, all contestants had the opportunity to submit a conference paper to the JSM proceedings, which is a benefit of participating in the *Joint Statistical Meetings* (https://www.amstat.org/ASA/Meetings/Joint-Statistical-Meetings.aspx). We briefly describe some of the analyses that were published in the JSM proceedings.

Vishnyakova (2016) examined the effect of motor vehicle crashes with young children as passengers. As we know, child restraints for cars, like car seats, can reduce injuries and the risk of child fatalities. This author combined data from the GES with demographic data from the *American Community Survey* to examine potential associations between families with low education and income and whether or not child restraints are used. The study showed that there is an association, and using multiple logistic regression, she found that the odds of not using restraints were higher for geographic areas with the highest socioeconomic deprivation.

Jadoo (2016) conducted a study that sought to find areas in the United States, which seemed to be the deadliest locations for drivers. Additionally, he looked at what factors could be used to predict accidents, which would subsequently highlight the factors contributing to the accidents. Interestingly, he also looked at the effects of high population areas and what sub-populations would be most affected by traffic accidents.

Heiberger (2016) took a visualization approach to explore the data, which is in keeping with the original focus of the *Data Challenge*. He looked at factors associated with car accidents where the drivers are teenagers. His main research question asked: "Does the tendency of teenage drivers to be involved in automobile accidents increase dramatically with the number of passengers in the car?" Answering this question might lead to policies limiting the number of passengers with teenage

**Fig. 1** Pictured left to right are Brennan Bean (Student Chair of the Utah State University Data Analytics Club, now an Assistant Professor at USU's Department of Mathematics and Statistics) and Wendy Martinez (one of the Guest Editors of this Special Issue)

drivers. His visualizations seem to indicate that teenager drivers do indeed drive less safely with more passengers in the car.

## 5 Supplementary materials

This special issue followed the precedent set by the 2013 *Data Expo* (Hofmann et al. 2019) for reproducibility, and we asked authors to upload supplementary materials to a Github site: https://github.com/asa-stat-computing-and-graphics/COST-DataExpo-2016. This included any files (Tex, bib, figures), data sets, and computer code (project files, macros, PROCS, R files, Shiny apps, etc.) they created as part of their analysis and reporting. It is particularly important that the computer code be reviewed prior to publication, and we were fortunate to have the *Utah State University* (USU) *Data Analytics Club* volunteer to review the code (Bean 2019). The *USU Data Analytics Club* is a student chapter of the *American Statistical Association* (see https://www.amstat.org/ASA/Membership/Student-Chapters.aspx for information on ASA student chapters). This was a great opportunity for students to serve as reviewers and to learn about reproducible research. See Figs. 1 and 2 for students' and one of the guest editors' hard at work reviewing the code.

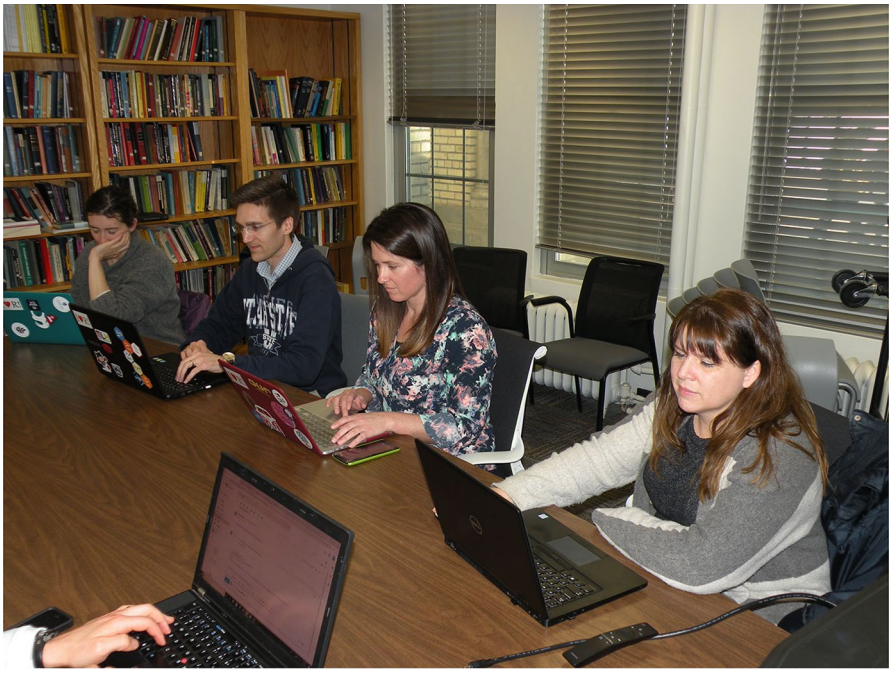

**Fig. 2** Evaluators of the supplementary materials from left to right are Angie Merritt, Eric McKinney, Kristi Reutzel, and Jill Lundell

**Acknowledgments** The Guest Editors and authors would like to thank the Editors, the referees, and the journal management staff of *Computational Statistics* and Springer for all of their help and patience as we worked through the process of preparing this special issue. A special thank you goes to Jürgen Symanzik (Past Editor-in-Chief) as he helped us get approval for this issue and helped shepherd us through the process. We also thank Lucy D'Agostino and Samantha Tyner for their help setting up the Github site. We are grateful for the support of the sponsoring sections of the ASA and our judges Nancy Bates, Stephen Brumbaugh, Morgan Earp, Jennifer Parker, and Juergen Symanzik. Finally, a special thank you to Spyros Bakas from Springer who supported color printing at no charge and enabled this fully reproducible special issue.

## References

Auerbach J, Eshleman C, Trangucci R (2021) A hierarchical Bayes approach to adjust for selection bias in before-after analyses of Vision Zero policies. Comput Stat. https://doi.org/10.1007/s00180-021-01070-x

Bean B (2019) USU data analytics club reviews code submissions with ASA president-elect. AMSTAT News. https://magazine.amstat.org/blog/2019/04/01/usu_april2019/

Cook D (2014) The 2011 data expo of the American Statistical Association. Comput Stat 29:117–119. https://doi.org/10.1007/s00180-013-0474-x

Coyle P, Chen C, Dabbish N (2016) Analysis of reported drowsy driving exploring subpopulation risk with weighted contingency table tools. JSM Proceedings. American Statistical Association, Alexandria, VA, pp 3018–3029

Coyle P, Chen C, Dabbish N (2021) Analysis of drowsy driving. Comput Stat. https://doi.org/10.1007/s00180-021-01071-w

Gunes A, Farrow R, Liu H, Moore C, Ng HKT, Stokes L, Xu Y, Xu Z, Yan Y, Zhong Y (2021) Predictive modeling of maximum injury severity and potential economic cost in a car accident based on the General Estimates Systems data. Comput Stat. https://doi.org/10.1007/s00180-021-01074-7

Heiberger RM (2016) Visualizing crash data by age group, time, and year. JSM Proceedings. American Statistical Association, Alexandria, VA, pp 2978–2981

Hofmann H, Wickham H, Cook D (2019) The 2013 data expo of the american statistical association. Comput Stat 34:1443–1447. https://doi.org/10.1007/s00180-019-00923-w

Jadoo M (2016) U. S. Roadways. In: JSM Proceedings. American Statistical Association, Alexandria, VA, pp 1820–1837

Murrell P (2010) The 2006 data expo of the American Statistical Association. Comput Stat 25:551–554. https://doi.org/10.1007/s00180-010-0207-3

Ng T, Stokes L, Zhong Y, Farrow R, Moore C, Alkan G, Liu H, Xu Z, Xu Y, Yan Y (2020) Predicting the potential economic cost of a car accident under different circumstances. Comput Stat.

Philips C, Garrett R, Tatro A, Fisher T (2016) Crash-safety ratings and the true assessment of injuries by vehicle. JSM Proceedings. American Statistical Association, Alexandria, VA, pp 3937–3946

Philips C, Garrett R, Tatro A, Fisher T (2021) An analysis of crash-safety ratings and the true assessment of injuries by vehicle. Comput Stat. https://doi.org/10.1007/s00180-021-01072-9

Roy D, Deshpande V, Linder MH (2016) A cluster-based taxonomy of bus crashes in the United States. JSM Proceedings. American Statistical Association, Alexandria, VA, pp 2731–2755

Roy D, Deshpande V, Linder MH (2021) A cluster-based taxonomy of bus crashes in the United States. Comput Stat. https://doi.org/10.1007/s00180-021-01073-8

Vishnyakova A (2016) Evaluating the use of child restraint systems and resulting injury and fatalities using demographic and social characteristics of driver's home zip code. JSM Proceedings. American Statistical Association, Alexandria, VA, pp 2021–2035