Datasheet for 'Age and Gender Dynamics in Traffic Accidents: A 2019 Canadian Case Studyt'*

Sinan Ma

April 19, 2024

This study analyzes traffic collision rates in Canada by examining the 2019 National Collision Database and Driver's License Statistics, employing a Bayesian regression model. Results show that drivers aged 16-24 have the highest accident rates, and males are involved in collisions more frequently than females. These insights emphasize the need for educational programs and policy adjustments that specifically address young and male drivers to reduce their risk. By pinpointing these demographic risk factors, the research supports the development of effective safety measures, aiming to lower traffic-related injuries and deaths.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to enable analysis of traffic collision rates in Canada. There was a requirement for a resource that compiled demographic and incident data to facilitate advanced statistical modeling. To address this need, the dataset was created, merging information from the National Collision Database and Driver's License Statistics to provide a foundation for the study.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The **National Collision Database** curated by Transport Canada, and the **Permis de Conduire** data created by the Société de l'assurance automobile du Québec (SAAQ).

[*] Code and data are available at: https://github.com/Sinanma/Traffic_Collision_Analysis.git.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

   - Canada government.

4. *Any other comments?*

   - No

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

   - The instances in the dataset represent individual traffic collisions and licensed drivers. The collision instances include various environmental and demographic details, whereas the driver instances include demographic distribution data.

2. *How many instances are there in total (of each type, if appropriate)?*

   - 21 for **Permis de Conduire** data. 22 for **National Collision Database** data.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset are contain all possible instances for a larger set.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Both datasets were included a lots of Raw data. C_YEAR, C_MNTH, C_WDAY, C_HOUR, and etc..

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Yes, the data are target associated with the real collision data and driver's licence data.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- Yes, the reseaon might be the privacy concerns, or the data collection limited, such as driver's licence number.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - Yes, the relationships between specific instances in the dataset are made clear. For instance, age, gender, and weather condition can all influence the collision severity.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - No, we thinking the datasets from opendatacanada are very professional.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - No, we thinking the datasets from opendatacanada are very professional.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - While the OpenDataCanada dataset on National Collision Database in 2019 is mainly self-contained, it may make use of other sources for further context or analysis, including academic articles, government reports, and other datasets. These external resources might not be included in official archive versions that contain the external resources as they were at the time the dataset was developed, and there are no guarantees that they will stay consistent over time. Depending on where they came from, external resources may have their own limitations, such as fees or licence requirements. One should consult the sources directly or the dataset's description for information on access and any related limitations.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - OpenDataCanada's datasets is unlikely to contain any confidential information. It is intended for public use while adhering to privacy rules, which means that any personally identifying information is anonymized or aggregated to safeguard individual privacy. This approach assures that the dataset is suitable for study and analysis while maintaining personal anonymity.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - OpenDataCanada's datasets is unlikely to contain anything that is immediately harmful, insulting, threatening, or anxiety-inducing, as it focuses on statistical statistics on shelter use and capacity. It is intended for objective research and policy-making, yet the subject of homelessness may elicit strong emotional responses due to the systemic challenges it raises.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - OpenDataCanada's datasets for National Collision Database and Permis de Conduire defines sub-populations based on gender, age, severity, weather condition and other variables, and displays their distribution as counts or percentages.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - OpenDataCanada's datasets is intended to ensure that no people can be identified, either directly or indirectly. It uses data anonymization and aggregation to protect privacy and comply with data protection rules, ensuring that no personal identifiers or sensitive information are included.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The OpenDataCanada's datasets may include demographic information such as age, gender, and perhaps race or ethnic origins, in order to analyse service utilisation and needs. However, it is unlikely to include sensitive information such as sexual orientation, religious beliefs, political ideas, financial, health, biometric, or genetic data, official identification, or criminal past. Any contained data is handled using privacy safeguards to prevent individual identification and to ensure compliance with data protection legislation.

16. *Any other comments?*

    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

   - The OpenDataCanada's datasets data collection involves gathering reports from police and transportation agencies, including various environmental and demographic factors related to each incident. Validation methods described in research techniques would be used on any derived or inferred data to make sure it is correct and reliable. Because the dataset is used to plan and make policies for public services, keeping its quality and safety is very important.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

   - The OpenDataCanada's datasets was put together by police or government agent entering data by hand and using software to handle the data. Validation includes making sure the data is correct and doing regular audits while following strong quality and moral standards to get accurate data.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

   - The OpenDataCanada's datasets is meant to be complete, with all the important data included. This dataset doesn't usually use sampling methods because it's meant to give a full picture that can help with policy and service planning.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - Police, government and transportation agencies.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - Every year.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- Since the penDataCanada's datasets is made up of administrative data that is normally collected by the Canadian government, it's possible that it didn't go through a separate ethical review process by an institutional review board.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - I collect the data from the websites.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - The question notified about the data collection should be necessary.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - The people's consent should be obtained before any data is collected.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - No

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No

12. *Any other comments?*

    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

- The study analyzed collision data and demographic data from the Canada's open data platform using R and tools from the tidyverse. Data pre-processing, cleaning, and labeling were performed using various R packages, such as dplyr, readr, model summary, janitor, tibble, and ggplot2, and etc..

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

   - The study used a dataset to analyze collision happen in Canada in 2019, focusing on relationship between demographic variable with collision. A Bayesian analysis model was used to investigate occupancy patterns and service utilization.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

   - Git hub: https://github.com/Sinanma/Traffic_Collision_Analysis.git

4. *Any other comments?*

   - No

Uses

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - The document describes an open access dataset that can be found in a public GitHub repository. It is meant to be shared with policymakers, researchers, and other people who are interested. This method encourages people to work together, repeat experiments, and do new analyses. It also stresses ethical, responsible data use, and privacy issues. Users are asked to follow the rules and terms of service.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The dataset is distributed via GitHub, a platform for easy access, sharing, and collaboration on code and datasets. It supports version control, issue tracking, and discussion among users. No DOI.

3. *When will the dataset be distributed?*

   - The document states that the dataset is available on GitHub, accessible to the public as of the study's publication. Users can access the dataset by visiting the provided link, but should check the repository for updates and changes, as open-source projects can be dynamic. Reviewing the commit history or release section can provide insights into its development.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - Users should visit the relevant GitHub repository to review licensing information, such as attribution and restrictions. They are encouraged to consult the dataset's GitHub repository for the most up-to-date licensing terms, which will ensure ethical and legal use.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

   - No

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

   - Users should visit the relevant GitHub repository to review licensing information, such as attribution and restrictions. They are encouraged to consult the dataset's GitHub repository for the most up-to-date licensing terms, which will ensure ethical and legal use.

7. *Any other comments?*

   - No

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - The document describes an open access dataset that can be found in a public GitHub repository. It is meant to be shared with policymakers, researchers, and other people who are interested. This method encourages people to work together, repeat experiments, and do new analyses. It also stresses ethical, responsible data use, and privacy issues. Users are asked to follow the rules and terms of service so they can contribute positively to the ongoing conversation about homelessness and stable housing.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- The dataset is distributed via GitHub, a platform for easy access, sharing, and collaboration on code and datasets. It supports version control, issue tracking, and discussion among users. No DOI.

3. *When will the dataset be distributed?*

   - The document states that the dataset is available on GitHub, accessible to the public as of the study's publication. Users can access the dataset by visiting the provided link, but should check the repository for updates and changes, as open-source projects can be dynamic. Reviewing the commit history or release section can provide insights into its development.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - Users should visit the relevant GitHub repository to review licensing information, such as attribution and restrictions. They are encouraged to consult the dataset's GitHub repository for the most up-to-date licensing terms, which will ensure ethical and legal use.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

   - No

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

   - Users should visit the relevant GitHub repository to review licensing information, such as attribution and restrictions. They are encouraged to consult the dataset's GitHub repository for the most up-to-date licensing terms, which will ensure ethical and legal use.

7. *Any other comments?*

   - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - OpenDataCanada platform.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - donnees.ouvertes@saaq.gouv.qc.ca and TC.Open-Ouvert.TC@tc.gc.ca

3. *Is there an erratum? If so, please provide a link or other access point.*

    - No

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - It can be made by creators at any time. Communication about updates could be facilitated through commit messages, GitHub's release feature, issues, discussions, or README file updates. Users are encouraged to follow the repository for updates.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - No

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - GitHub allows for archival and access of older versions through commit history and releases, but dataset creators are responsible for actively supporting or updating them. If support ceases, communication may occur through README files, releases announcements, or issues discussions. Users should stay engaged with the repository for updates.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - Email author, contact information avaibile on github.

8. *Any other comments?*

    - No

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.