Age and Gender Dynamics in Traffic Accidents: A 2019 Canadian Case Study*

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.

Table of contents

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

1 Introduction

Road traffic injuries are a public health issue, claiming approximately 1.19 million lives globally every year and being the leading cause cause of death among children and young adults aged 5–29 years. Highlighted by the World Health Organization, these statistics illustrate the critical need for further research and effective policy making in traffic safety (World Health Organization 2023). In Canada, the complexity of traffic incidents is detailed within the 2019 national collision dataset, which provides insights into the demographic factors-age, gender-that influence the severity of these events.

This study is focusing on the impact of demographic factors such as age and gender on traffic collision rates. Utilizing a Bayesian statistical model to provide accurate estimates of these effects, implemented through R—a comprehensive environment for statistical computing and graphics (R Core Team 2023). The methodology and analytical strategies employed are learned from *Telling Stories with Data* by Rohan Alexander (Alexander 2023), which provides the statistical framework and R code that underpin our analysis.

The estimand of this paper is the effect of these demographic variables—age and gender—on the severity of traffic incidents. By focusing exclusively on these factors, this research aims to isolate and quantify their specific impacts, offering detail into how demographic characteristics modulate collision outcomes.

The findings reveal variances in accident rates by age and gender: younger drivers (aged 16-24) exhibit the highest accident rates, while older drivers (65 and above) show the lowest, which suggests differing risk levels associated with driver experience. Additionally, male drivers are found to have higher accident rates than female drivers, indicating potentially riskier driving behaviors or greater exposure to traffic due to occupational demands.

This study provides insights for traffic safety regulators, urban planners, and policymakers by detailed understandings of the demographic factors that significantly influence traffic accidents, which is essential for developing targeted interventions designed to reduce traffic-related injuries and fatalities.

The paper begins by introducing the broader context and motivation behind our study. Section 2 introduces the dataset utilized for analysis, explores the dataset and key variables, supported by visualization. Section 3 describes the use of the Bayesian regression model. Section 4 presents the findings, offering detailed interpretations of the data. Section 5 discusses the implications, study limitations, and directions for future research.

2 Data

This paper uses two datasets from Open Canada's data portal. The first is the **National Collision Database** which provides records of traffic collisions across Canada, including various environmental and demographic factors related to each incident. The complete dataset for 2019 is available at National Collision Database (Transport Canada 2019b), with further details provided on its main overview page (Transport Canada 2019a).

The second dataset, **Permis de Conduire**, offers demographic distribution of drivers. This can be accessed through Permis de Conduire 2019 (Société de l'assurance automobile du Québec (SAAQ) 2019) and further explored on its overview page (Transport Canada 2019c).

Data was collected, analyzed, tested, and visualized using R statistical programming software (R Core Team 2023), with packages: `tidyverse` (Wickham et al. 2019), `rstanarm` (Goodrich et al. 2022) `knitr` (Xie 2020), `here` (Müller 2020), `janitor` (Firke 2023), `lubridate` (Grolemund and Wickham 2011), `arrow` (Richardson et al. 2024), `dply` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2024), `ggplot2` (Wickham 2016), `testthat` (Wickham 2011), and `scales` (Wickham, Pedersen, and Seidel 2023).

2.1 Dataset Description

The **National Collision Database** curated by Transport Canada, captures details from over 200,000 traffic incidents, including driver age, gender, and collision severity. Data collection involves gathering reports from police and transportation agencies, including various environmental and demographic factors related to each incident. Merged with the **Permis de Conduire** data from the Société de l'assurance automobile du Québec (SAAQ), which offers demographic distribution of licensed drivers, including age and gender. The combined datasets enable a clear analysis of factors affecting traffic collision severity in Canada.

This translates the incidents and demographic details into analyzable data, shows the influence of age and gender on collision outcomes. This helps us better identify patterns and risk factors, thus provides insight for the development of informed traffic safety strategies.

2.2 Measurement and Variables

Age: Numerical then categorized into predefined age groups (e.g., 16-19, 20-24, etc.); the age of the driver involved in the collision.

Gender: Categorical; the gender of the driver, categorized as “Male”, “Female”.

Collision Severity: Categorical; indicates the severity of the collision, categorized as “Fatal” or “Non-Fatal”.

Year: Categorical; the year in which the collision occurred, we focus on the data for 2019.

Accident Rate per 1000 Drivers: Numerical; calculated as the number of accidents per age and gender group per 1000 drivers, providing a normalized measure of collision risk.

2.3 Data Cleaning and Preparation

Data was cleaned and prepared for analysis, which involved handling missing values, and correcting data entry errors. The cleaned dataset was then used for all subsequent analyses, and these steps guarantee the validity of the analysis.

The collision rate per 1000 drivers was calculated by merging the data on traffic incidents, classified by age and gender, with the data on the number of licensed drivers in those categories. The number of incidents for each group was divided by the number of drivers, and this ratio was then multiplied by 1000 to establish the collision rate per thousand. The processed data was saved to help with comparison across different demographic groups.

2.4 Graphical and Statistical Analysis

Graphical representations and statistical analyses were used to illuminate the data’s properties, exploring the interplay between variables and traffic collision outcomes.

Accident Rate per 1000 Drivers by Age Group (Table 1) presents the relationship between traffic accidents and the number of licensed drivers, sorted by age, indicating specific age groups with heightened collision rates.

Accident Rate per 1000 Drivers by Gender (Table 2) contrasts accident rates between genders (male and female), suggesting a higher incidence among male drivers and indicating possible behavioral influences.

Distribution of Accidents by Driver Age (Figure 1) illustrates the variation in accident rates across age groups, highlighting the elevated risk among younger drivers and the need for focused safety measures.

Distribution of Accidents by Driver gender (Figure 2) illustrates the differences in accident rates between male and female drivers, bringing attention to potential areas for targeted safety initiatives.

These visualizations provides better understanding in the data and assist in making conclusions about traffic safety and accident prevention strategies.

Table 1: Accident Rate per 1000 Drivers by Age Group

Age Group	Number of Accidents	Number of Licensed Drivers	Accident Rate per 1000 Drivers
16-19	11,655	210,120	55.47
20-24	20,038	360,215	55.63
25-34	36,925	885,713	41.69
35-44	32,369	998,515	32.42
45-54	29,643	957,857	30.95
55-64	24,842	1,095,137	22.68
65 or older	21,432	1,192,420	17.97

The table "Accident Rate per 1000 Drivers by Age Group" (Table 1) shows that drivers aged 16-19 and 20-24 have the highest accident rates per 1000 drivers, at 55.47 and 55.63, respectively. This indicates these younger drivers are more prone to accidents, likely due to inexperience. In contrast, older drivers (65 and above) have the lowest rate at 17.97, reflecting more conservative driving habits and greater experience. These statistics highlight the need for targeted educational initiatives to reduce risks for younger drivers.

Table 2: Accident Rate per 1000 Drivers by Gender

Gender	Number of Accidents	Number of Licensed Drivers	Accident Rate per 1000 Drivers
Female	71,132	2,777,577	25.61
Male	106,783	2,929,083	36.46

The table "Accident Rate per 1000 Drivers by Gender" (Table 2) reveals higher accident rates among male drivers (36.46) compared to female drivers (25.61). This disparity could be due to males engaging in riskier driving behaviors, possibly influenced by societal norms, or having higher annual mileage, which increases accident exposure. Occupational demands and differences in safety attitudes or vehicle types might also contribute to this trend.

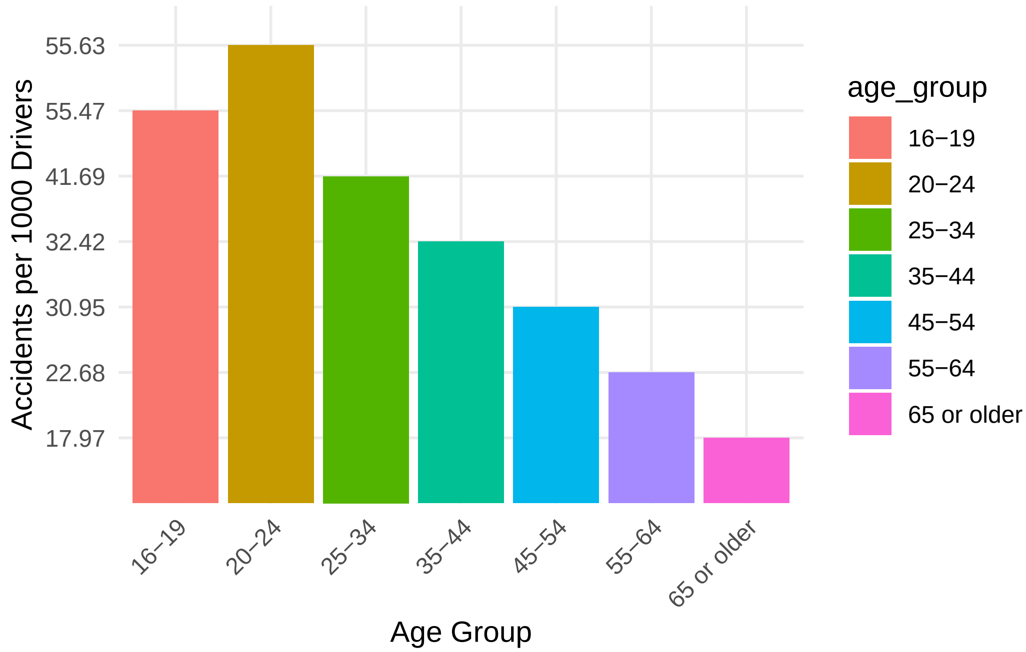

Figure 1: Distribution of Accidents by Driver Age

The bar graph "Distribution of Accidents by Driver Age" (Figure 1) illustrates accident rates per 1000 drivers across different age groups. The tallest bars for the youngest groups confirm their high accident rates, while shorter bars for older age groups depict lower rates, suggesting safer driving habits develop with age. This graph effectively highlights key risk groups and intervention points.

Figure 2: Distribution of Accidents by Driver gender

The bar graph "Distribution of Accidents by Driver Gender" (Figure 2) shows accident rates for male and female drivers. The pronounced difference between the two bars highlights the higher risk among male drivers, providing a clear visual of the gender disparity in traffic accidents.

3 Model

In this study, we employed a Bayesian statistical tool to address two key aspects of traffic safety analysis using data from the 2019 National Collision Database and Driver's License Statistics:

1. **Severity of Traffic Collisions**: Predicting outcomes of traffic collisions based on age and gender.
2. **Demographic Influences on Driver Counts**: Examining the effects of demographic factors on the frequency of drivers involved in accidents.

The Bayesian approach was chosen due to its flexibility in incorporating prior knowledge and its efficacy in handling uncertainty, which is particularly beneficial in analyzing the multifaceted nature of traffic incidents.

3.1 Model Setup

3.1.1 Collision Severity Model

We modeled the severity of traffic collisions using a logistic regression, presented in a hierarchical Bayesian format:

$$y_i|\mu_i \sim \text{Bernoulli}(p_i)) \tag{1}$$
$$\text{logit}(p_i) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Gender} \tag{2}$$
$$\beta_0, \beta_1, \beta_2 \sim \text{Normal}(0, 2.5) \tag{3}$$

y_i: Indicates whether a collision is fatal (1) or not (0) for each incident i.

p_i: Probability of the collision being fatal, modeled as a function of age and gender.

β_0(Intercept): Baseline log-odds of fatality when age and gender are at their reference levels.

β_1, β_2: Coefficients for age and gender, respectively, showing their influence on the log-odds of fatality.

Priors: The priors for the coefficients (β_0, β_1, β_2) are normally distributed around 0 with a standard deviation of 2.5. This prior setup implies no initial effect (bias) is assumed, and a moderate degree of uncertainty allows the data to significantly inform the parameter estimates.

3.1.2 Demographic Factors Model

Driver counts, influenced by demographic variables, are modeled using a Poisson regression framework as detailed below:

$$z_i|\lambda_i \sim \text{Poisson}(\lambda_i) \tag{4}$$
$$\log(\lambda_i) = \gamma_0 + \gamma_1 \cdot \text{Age\_Group} + \gamma_2 \cdot \text{Gender} \tag{5}$$
$$\gamma_0 \sim \text{Normal}(0, 2.5) \tag{6}$$
$$\gamma_1, \gamma_2 \sim \text{Normal}(0, 2.5) \tag{7}$$

z_i: Count of drivers involved in collisions for demographic group i.

λ_i: Rate parameter for the Poisson distribution, modeled logarithmically to relate the linear combination of predictors to the expected count.

γ_0: The intercept, indicating the baseline rate of driver counts when all predictors (age and gender) are at their reference levels.

γ_1, γ_2: Regression coefficients for the categorical predictors Age_Group and Gender, respectively.

Priors: The priors for the intercept (γ_0) and coefficients (γ_1, γ_2) are set as normal distributions centered at zero with a standard deviation of 2.5. This choice reflects no initial bias towards any outcome, with a moderate degree of uncertainty that allows the data to inform the final model outcomes effectively.

3.2 Model Justification

The use of Bayesian models in this study provides advantages for analyzing traffic safety data, particularly their capability to incorporate prior knowledge and effectively handle uncertainty. Bayesian statistics integrate prior distributions, allowing the inclusion of past research into the analysis, thus enhancing the strength and reliability of the findings. This integration is valuable in traffic safety studies where historical data can guide current model assumptions.

Bayesian inference quantifies the uncertainty inherent in traffic data analysis, offering estimates of model parameters along with intervals. Additionally, the flexibility of Bayesian models accommodates complex relationships within traffic safety data through hierarchical structures.

For modeling the severity of traffic collisions, logistic regression is used due to the binary nature of the outcome—fatal versus non-fatal. Logistic regression is well-suited for providing probabilities of these outcomes, crucial for evaluating risk factors. The hierarchical setup of this model allows for nuanced examination of age and gender impacts on collision severity, accounting for variability across different demographic groups.

The influence of demographic factors on the frequency of drivers involved in accidents is modeled using Poisson regression. This approach is appropriate for count data and ensures that predictions are positive, aligning with the count nature of the data. The logarithmic link function used in Poisson regression is standard for maintaining the non-negativity of count predictions.

Both models improve the study's predictive accuracy regarding traffic accidents. These models provide insights into how demographic factors affect traffic safety, offering information for policymakers. The findings inform targeted safety measures and regulatory adjustments aimed at reducing traffic-related injuries and fatalities.

4 Results

4.1 Model Summaries

In this section, we present the findings from our statistical analysis of the relationship between age, gender, and the likelihood of traffic collisions. This analysis involves two models: a logistic regression model for collision severity and a Poisson regression model for demographic analysis.

4.1.1 Collision Severity Model

The logistic regression model (Table 3) assessing traffic collision severity provided estimates of the effects of age and gender. The model estimated a negative effect for the intercept, suggesting a baseline tendency towards non-fatal outcomes. A slight positive estimate for age indicates an increase in the odds of fatal collisions as age increases. Gender, coded as male, showed a positive effect, suggesting males are at a higher risk for fatal collisions compared to females.

4.1.2 Demographic Factors Model

The Poisson regression model (Table 3) analyzing demographic factors on the rate of driver involvement in collisions showed significant age and gender effects. The intercept was notably high, with younger age groups such as 20-24 and 25-34 showing a reduced rate of involvement when compared to the reference group. Age groups 65 or older showed a further reduced rate, indicating a decline in collision involvement with increasing age. Male gender showed an increased rate of involvement in collisions.

4.1.3 Model Fit and Diagnostic Metrics

The number of observations for the collision severity model was substantial, providing a robust dataset for analysis. The log-likelihood and other model fit indices such as ELPD and WAIC indicate the models' good fit to the data.

The results from the models highlight critical areas for intervention. The increased risk associated with younger drivers points to the potential benefit of targeted education and policy measures to improve driving safety within this demographic. The gender disparity suggests a need for further research into behavioral factors that may contribute to higher collision rates among males.

Table 3: Combined Model Summary: Traffic Collisions and Demographic Factors

	Traffic Collision Severity Model	Demographic Factors Model
(Intercept)	-5.29	11.54
	(0.07)	(0.00)
Age	0.01	
	(0.00)	
GenderMale	0.85	0.05
	(0.05)	(0.00)
Age_Group20-24		0.54
		(0.00)
Age_Group25-34		1.44
		(0.00)
Age_Group35-44		1.56
		(0.00)
Age_Group45-54		1.52
		(0.00)
Age_Group55-64		1.65
		(0.00)
Age_Group65 or older		1.74
		(0.00)
Age_Groupunder 16		-3.45
		(0.01)
Num.Obs.	177915	16
Log.Lik.		-1433.118
ELPD	-13137.2	-1804.1
ELPD s.e.	208.7	507.3
LOOIC	26274.3	3608.2
LOOIC s.e.	417.4	1014.6
WAIC	26274.3	5279.7
RMSE	0.12	7867.77

4.2 Distribution of Fatal Accidents by Age Group

The bar chart below illustrates the rate of fatal accidents per 1000 drivers segmented by age group, which aligns with the quantitative findings from our regression analyses.

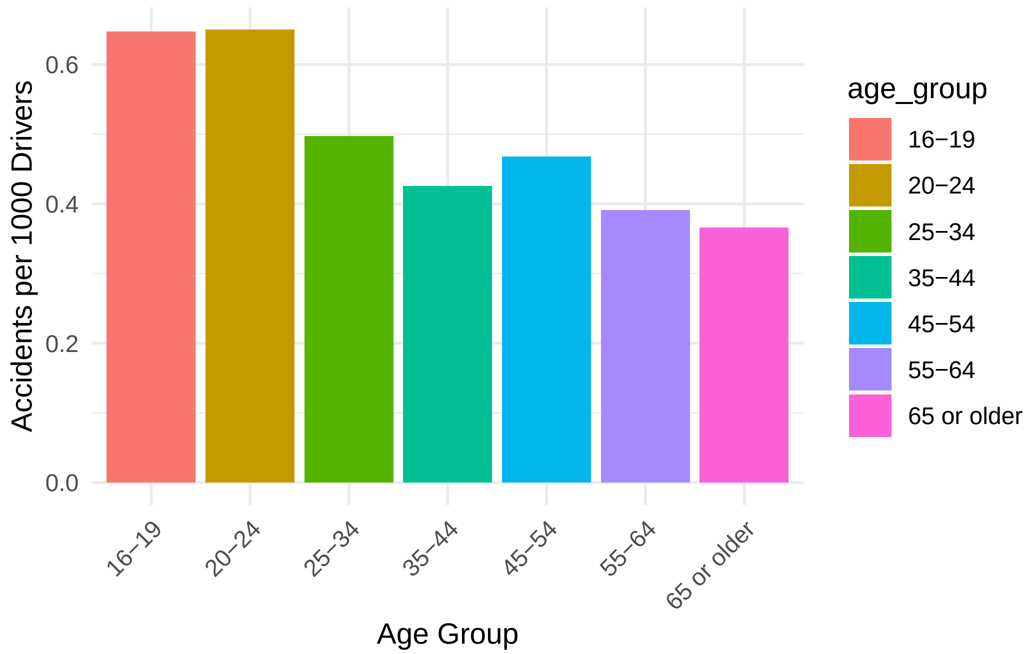

Figure 3: Distribution of Fatal Accidents by Driver Age

Figure 3 presents a distinct pattern: the highest rates of fatal accidents are amongst the youngest drivers, with a pronounced decrease as age increases. The 16-19 age group exhibits the highest accident rate, which sharply contrasts with the notably lower rate observed in the 65 or older group. This visual representation further substantiates the model's indication that younger drivers are at an elevated risk of fatal accidents, potentially due to less experience or riskier driving behaviors.

The consistency in the decline of accident rates across age groups from this graph is particularly informative. It indicates that drivers aged 65 or older are the least likely to be involved in fatal accidents, which may reflect more cautious driving habits or reduced driving exposure.

The comprehensive analysis of collision data through our models, supported by the visual evidence from the age distribution graph, illustrate the critical role of age in traffic safety. These findings provide a compelling argument for tailored preventive measures focusing on younger drivers, aiming to reduce the higher rates of fatal accidents within this demographic. Developing strategies that address the specific needs and risks associated with different age groups can contribute significantly to enhancing overall road safety.

5 Discussion

5.1 Insights from Traffic Collision Analysis in Canada

5.1.1 Age and Gender Differences

This study demonstrates the significant role that age plays in traffic collisions, with young drivers (aged 16-24) showing the highest rates of incidents. This trend highlights the need for targeted interventions such as specialized training and awareness programs aimed at young drivers to enhance their understanding of road safety.

Drivers aged 65 and above have lower accident rates, potentially due to more cautious driving habits and decreased frequency of driving. These findings could guide policies to maintain and possibly extend driving safety among older adults.

In terms of gender, the analysis reveals that male drivers are more frequently involved in accidents than female drivers. This difference suggests that male drivers might engage in more aggressive driving behaviors or have higher exposure to driving, which includes occupational driving that often involves higher risks.

5.1.2 Seasonal and Regulatory Influences

Exploring the influence of external factors like weather conditions and regulatory changes could provide further insights. The analysis of accident rates over multiple years could reveal the effects of new traffic laws or advancements in vehicle safety technology.

5.2 Limitations

5.2.1 Data and Scope Limitations

The primary dataset from 2019 offers a snapshot but lacks the scope to examine changes over time or the effects of recent global disruptions like the pandemic. Future studies should include more extensive temporal data and consider other influential factors such as socioeconomic variables, which might affect driving patterns and collision rates.

5.2.2 Analytical Constraints

The Bayesian regression approach, while sophisticated, depends heavily on the assumptions made about prior distributions and model parameters. Exploring a range of statistical methods and incorporating machine learning could potentially provide a more nuanced understanding of the factors that contribute to traffic accidents.

5.3 Future Research Directions

5.3.1 Broader Data Collection

Incorporating a wider array of data, including multi-year trends, socio-economic factors, and environmental conditions, would allow for a more detailed analysis of collision causes. This approach would enable researchers to identify and quantify the impacts of various risk factors more accurately.

5.3.2 Enhanced Analytical Techniques

Utilizing advanced analytics, including machine learning models, could help identify complex interactions between numerous risk factors. These techniques may reveal insights that are not apparent through traditional modeling approaches, such as hidden patterns or predictive indicators of high-risk areas.

5.3.3 Policy Implementation

Findings from this study suggest the need for differentiated driving education that addresses the specific risks associated with each demographic group, particularly young and male drivers. Additionally, regular reassessment of traffic laws and safety protocols is essential to adapt to new data and changing conditions effectively.

5.4 Conclusion

This research provides insights into the demographic factors that influence traffic collision rates in Canada, particularly highlighting the risks associated with young and male drivers. While the current study offers guidance for safety interventions, the identified limitations and suggested future research indicate the ongoing need to refine data collection and analytical methods. By expanding the scope of data and embracing new analytical technologies, stakeholders can better understand and mitigate the complexities of road safety. This proactive approach helps developing more effective traffic policies and reducing accident rates.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Boca Raton: CRC Press. https://tellingstorieswithdata.com/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Société de l'assurance automobile du Québec (SAAQ). 2019. "Permis de Conduire 2019." https://dq-prd-bucket1.s3.ca-central-1.amazonaws.com/saaq/Permis_Conduire_2019.csv.

Transport Canada. 2019a. "2019 National Collision Database (NCDB)." Open Canada. https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a.

———. 2019b. "National Collision Database (NCDB) – 2019 Collision Data." Online. https://open.canada.ca/data/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a/resource/1e2a6aa0-c35f-4736-984a-71f2eac1c1dc/download/2019_dataset_en.csv.

———. 2019c. "Transport Canada Open Data." https://open.canada.ca/data/en/dataset/f2269058-2c30-483c-8a83-ad7862ae860a.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

World Health Organization. 2023. "Road Traffic Injuries." https://www.who.int/news-room/

fact-sheets/detail/road-traffic-injuries.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.