Taking matters into your own hands

Women are learning self defense

Colin Conant, Sara Hamdy, Jack Smith

October 19, 2020

GitHub Repo

Code and data supporting this analysis is available at: “https://github.com/Saraahamdy/STA304H1-PS2-Group-70.git (https://github.com/Saraahamdy/STA304H1-PS2-Group-70.git)”

Abstract

This report displays the analysis of the Canadian General Social Survey from 2014, which is focused on experiences of victimization Canadians might have experienced. Our model, which is a logistic regression model, is designed to help predict how likely certain demographics of people would be to sign up for self defense classes. We found that younger women were more likely to sign up for self defense classes the less safe they felt walking alone. Our evidence should be used to redirect authorities, support, and protection tools towards demographics that could be at risk.

Introduction

In our analysis we find out how safe people feel walking home at night and whether or not they took self defense classes. Our data is focused on types of victimization Canadians might have experienced. This includes data such as the general feeling of safety in one’s neighborhood, or experience of childhood abuse. We have chosen to analyze the chances of people joining self defense classes depending on specific categories. These categories consist of whether they are men or women, how safe they feel walking alone, and which age group they are in. We expect to see that younger women generally feel more inclined to sign up for self defense classes than men. Feeling unsafe walking alone will also push them to sign up for classes. We think people around their twenties and thirties will be taking classes more, since it is around this time people may begin to live alone. We will model our data using logistic regression, a model that displays the probability of binary response variables, yes or no answers. Our model will allow us to observe which groups of people can be targeted for self defense classes to grant them more comfort in their daily lives.

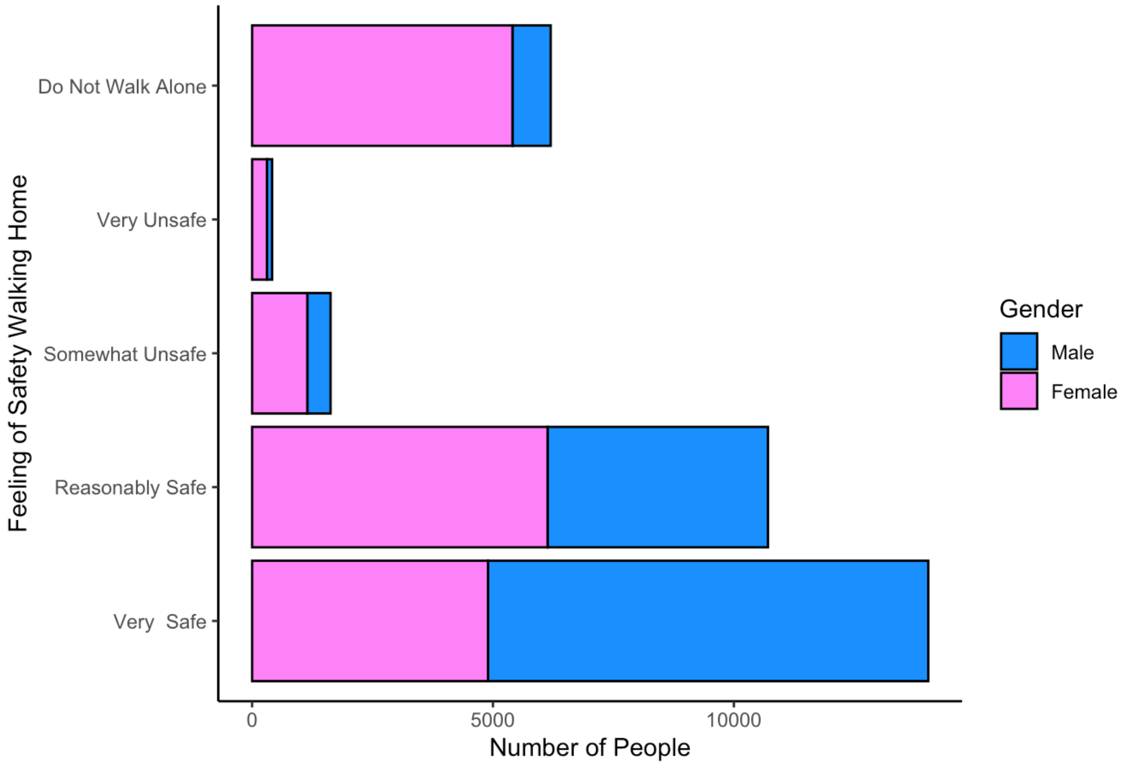
Data

The data we will be using is from the Canadian General Social Survey performed in 2014. The core focus of the survey was victimization. Some examples of the questions asked were “Abuse by a child, relative, friend or caregiver” or “Confidence in Police”. The data originally consists of 790 columns, the number of variables, and 33089 rows, the number of observations. After cleaning the data up we selected 60 of the 790 variables to further develop our analysis. The data is collected through a survey. It was distributed online as well as through phone interviews, where the phone numbers were randomly generated. The survey is designed to receive a summary on experience with violence. We chose this data to get a sense on how safe Canadians might feel, and if they take steps towards feeling safe in certain situations, like being alone. Wanting to know how Canadians feel overall, our target population is Canadian citizens. Our frame population are the individuals who come across the survey online or randomly dialed, and our sample population are the people who chose to answer the survey.

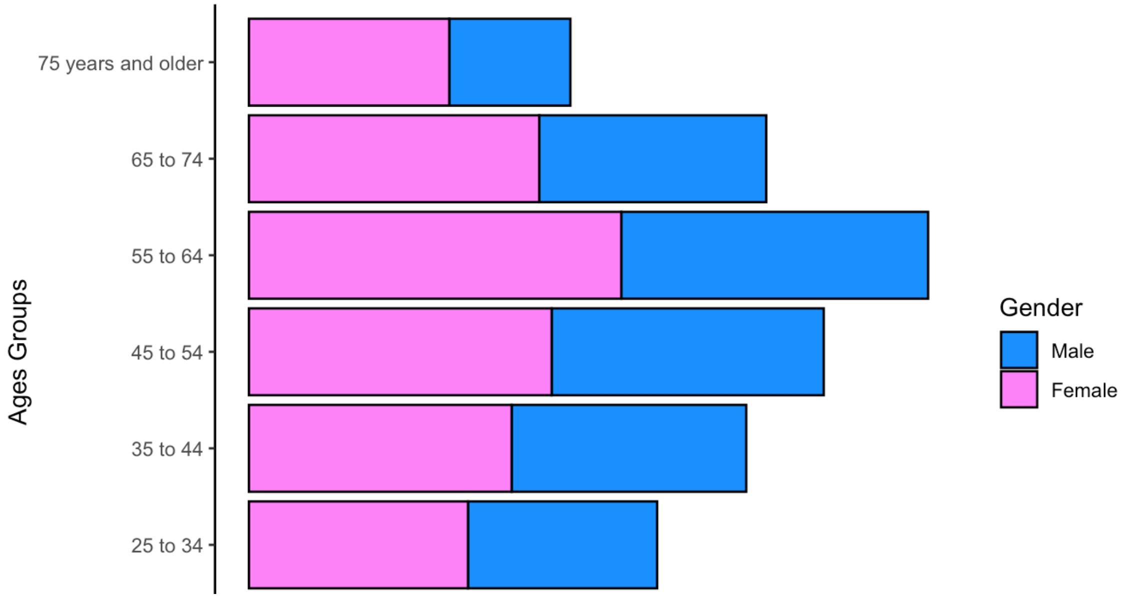
The data collected is very thorough, asking multiple variations of a question. For example, steps people took towards feeling safe had options such as, getting a dog, or taking self defense classes. This allowed us to have a deeper understanding of the individuals surveyed. While our data is thorough, we do have a handful of responses not available because of lack of response. Since it is a very small portion of our data, it should not greatly affect our study. Furthermore, most of our data’s variables are categorical. While this might not be seen as a major drawback, it does limit the variety of opinions we could have received, narrowing possible variation of the results accumulated. Since our variables are strictly categorical, the model used is limited to logistic regression. Lastly, our data is only collected from Canadians who are older than 15. This limits our demographic evaluations by excluding younger children who may have relevant data or experience for the survey.

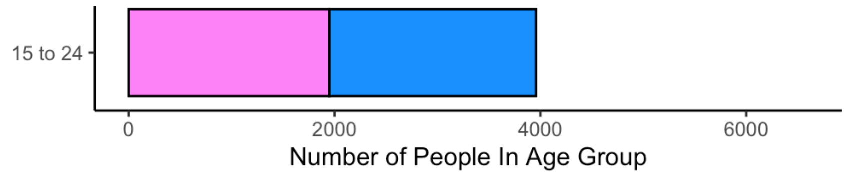
As seen below we have displayed the frequency of each age group with color coding representing men and women in each group. The bar graph shows that the most responses were recorded from the age group 55-64 and around 8000 responses from the ages 15-34. As well as a bar chart representing the frequency of how safe each category felt also displaying the amount of men and women in each group. This graph displays that the majority of the respondents felt very safe walking home and it had more male responses than men. Furthermore, the majority of respondents who did not walk alone were women. Men generally felt safer than women walking alone.

Feeling of Safety Walking Home Chart (Fig. 1)



Ages of Respondents Bar Chart (Fig. 2)





Model

We chose R as our software to do statistical analysis, utilizing various packages, including ggplot2, tidyverse, knitr, survey, broom, and car. These packages have access to functions that helped to create plots, make calculations, and create a model for the data. The function svyglm() from the survey package facilitated the creation of the selected model for analyzing the data, a binary multiple logistic regression utilizing a frequentist method. This model was selected because the response variable, whether or not a respondent has taken a self defense course, is a binary categorical variable, and the explanatory variables are categorical variables as well.

Age group was selected as one of the explanatory variables because we wanted to explore the possibility that age would change the odds or attitude towards a self defense course. The data did not have a continuous age variable, and so age group was the next best variable to assess age. The binary categorical variable, sex of the respondent, was also selected as an explanatory variable as we expected that there may be differences in how men and women view or take self defense courses. Feeling of safety when walking alone was chosen as an additional explanatory variable as we suspected that there would be a relationship between how safe a respondent feels walking home and taking a course that would ideally increase confidence in feelings of safety. The feeling of safety while walking home was a categorical variable with outcomes: Very Safe, Reasonably Safe, Somewhat Unsafe, Very Unsafe, Do Not Walk Alone. We only selected 3 explanatory variables because we wanted to avoid using too many explanatory variables, which could result in overfitting.

A binary categorical response variable and categorical explanatory variables indicate that a multiple logistic regression would be an appropriate model. A linear or multilinear model would not be as appropriate because of the binary categorical response variable. A bayesian model was not used, as we opted for a frequentist method. A frequentist method was adopted because GSS data is gathered frequently and has large values of n, allowing us to avoid bias and identify trends in the long run. The model utilizes a finite population correction based on simple random sampling. The simple random sampling technique was taken from the methodology in which the data was collected, and the N value was calculated as the number of Canadians that are older than 15 in 2014. This was gathered from the Stats Canada website and is appropriate in the 2014 GSS data because the survey covered the territories as well as provinces, so no adjustment must be made for provincial only data [Government of Canada, 2015][Government of Canada, 2019].

Several assumptions however, must be made when using a multiple logistic regression model. First, the observations must be independent. Second, the response variable must be binary. Third, we must assume the logit function has a linear relationship with the variables. Finally, the explanatory variables must have low multicollinearity. This final assumption can be verified by using variance inflation factor and generalized variance inflation factor. These metrics quantify the correlation between explanatory variables [Fox & Monette, 1992]. After calculating the GVIF values, all GVIF^(1/(2*Df)) values are less than the common benchmark we selected of square root of 5, giving evidence that the explanatory variables are not correlated with each other[Buteikis, 2018].

Below is the model equation. In this equation, the intercept is calculated by taking into account ‘default’ values that the respondent is male, has a “very safe” feeling of safety walking home, and is in the age group 15-24. Beta 1 is the coefficient applied if the respondent is female. Betas 2 through 5 refer to coefficients applied depending on each of the options for the respondents feeling of safety walking alone. Betas 6 through 11 are applied based on the age group the respondent falls under. The betas are all dummy variables, as the explanatory variables are categorical. Dummy variable coefficients are derived in relation to the ‘default’ value that is incorporated into the intercept. The probability of a given outcome increases with the increase of the log odds of the outcome. Using the model logit-linear equation, we can derive that this means positive, larger coefficients indicate increases in probability.

Equation 1: Model equation:

$$log(\frac{p}{1-p}) \ = \ \beta_0 + \beta_1 Female + \beta_2 ReasonablySafe + \beta_3 SomewhatUnsafe + \beta_4 VeryUnsafe + \beta_5 DoNotWalkAlone + \beta_6 Age25to35$$
$$+ \beta_7 Age35to44 + \beta_8 Age45to54 + \beta_9 Age55to64 + \beta_{10} Age65to74 + \beta_{11} 75andOlder$$

Results

Table 1

term	estimate	std.error	statistic	p.value
(Intercept)	-1.9055898	0.0524397	-36.3386813	0.0000000
R_sex	0.1438686	0.0391561	3.6742355	0.0002389
as.factor(feeling_of_safety_walking)2	0.0615199	0.0419434	1.4667347	0.1424578
as.factor(feeling_of_safety_walking)3	0.3052958	0.0771861	3.9553191	0.0000766
as.factor(feeling_of_safety_walking)4	0.4110947	0.1426688	2.8814625	0.0039609
as.factor(feeling_of_safety_walking)5	0.0529450	0.0581078	0.9111523	0.3622218
as.factor(R_age_group)25 to 34	-0.1200663	0.0653648	-1.8368656	0.0662387
as.factor(R_age_group)35 to 44	-0.0484624	0.0615864	-0.7869010	0.4313455
as.factor(R_age_group)45 to 54	-0.1824872	0.0610684	-2.9882410	0.0028080
as.factor(R_age_group)55 to 64	-0.4228361	0.0615181	-6.8733547	0.0000000
as.factor(R_age_group)65 to 74	-0.7712889	0.0721559	-10.6892083	0.0000000
as.factor(R_age_group)75 years and older	-1.2717683	0.1002740	-12.6829310	0.0000000

After developing the model, the equation became:

Equation 2:

$$log(\frac{p}{1-p}) \ = \ -1.90559 + 0.14387 Female + 0.06152 ReasonablySafe + 0.30530 SomewhatUnsafe + 0.41109 VeryUnsafe$$
$$+ 0.05295 DoNotWalkAlone - 0.12007 Age25to35 - 0.04846 Age35to44 - 0.18249 Age45to54$$
$$- 0.42284 Age55to64 - 0.77129 Age65to74 - 1.27177 Age75andOlder$$

Examples of the math from our model:

This is an example of a young female age 15-24 who feels very unsafe walking home:

$$log(\frac{p}{1-p}) \ = -1.90559 + 0.41109$$

$$log(\frac{p}{1-p}) \ = -1.35063$$

$$\frac{p}{1-p} = e^{-1.35063}$$

$$p = e^{-1.35063}(1-p)$$

$$p = e^{-1.35063} - e^{-1.35063}p$$

$$p + e^{-1.35063}p = e^{-1.35063}$$

$$p(1 + e^{-1.35063}) = e^{-1.35063}$$

$$p = 0.2057674$$

Therefore she has a 20.58% chance of having attended a self defense course.

Following the same steps as before we find the probability of a man aged 75+ who feels very unsafe:

$$log(\frac{p}{1-p}) \ = -1.90559 - 1.27177 => p = 0.04002665$$

or a 4.00% chance of having attended a self defense course.

Discussion

The slope parameters of the regression represent the change in log odds in the likelihood someone has attended a self defense course. For example, the estimate of the parameter for a respondent that is female has a value of 0.14387, therefore the log odds of attending a self defense course increases by 0.14387 from the base log odds, which is the intercept. The intercept represents the log odds of attending a self defense course for someone that is male, feels very safe walking home and in the age bracket of 15-24. Each slope coefficient (ie. B1, B2, … B11) represents the change in log odds of attending a self defense course compared to the base person in our model described by the intercept (seen in table 1). Positive values of the slope parameter mean a respondent is more likely to attend compared to the intercept. Negative values reduce the odds of a respondent attending a self defense course compared to the intercept. It is important to note in interpretation that negative values in the slope and intercept do not mean there is a negative probability of attending. Rather, we are observing the change in log odds and can only derive likelihood, which becomes positive, when solving the model equation as seen in the examples.

Analysing the results of the logistic regression estimates more closely, we can see that there is an increase in the likelihood of attending a self defense course if the respondent was female compared to male. There was also an increase in the log odds, and therefore likelihood, for people who felt less safe walking home compared to respondents that felt very safe walking home. Looking at the age group parameter estimates, the probability of attending a self defense course seems to decrease with age. People in the age group of 15-24 were the most likely age group to attend a self defense course while those in the age group of 75 and older had the lowest likelihood of attending a course.

We created the figures 1 and 2, as well as table 1 so we could better visualize the population that we were examining in our model. In terms of men and women there looks to be an even split. This is evidence that the sex distribution of our sample data is close to the true population percentage of men and women. In terms of age groups, the amount of people in each bracket is increasing until the 55-64 age range and then it starts to decrease. Comparing these values to the 2014 Stats Canada population statistics [Government of Canada, 2015], we see that there seems to be a relatively even amount of people in each age group, with a decrease at around 65. This could be a potential issue as the age group totals do not seem to match Stats Canada's data, and could indicate a non-representative sample. The chart comparing respondents' feelings of safety walking home shows that most respondents feel very safe, reasonably safe, or do not walk alone. One difference that can be seen when comparing people's safety when walking home is that there appears to be more women than men that either feel unsafe walking home or do not walk alone. To make sure that there were not high multicollinearity between these explanatory variables, earlier we computed the GVIF values and did not find large multicollinearity between the variables.

The p-value of the regression represents the probability that the slope of each variable was achieved by chance. The p-value can be interpreted as the benchmark for sufficient evidence for the null hypothesis, that claims that the estimate equals 0. We use a 5% benchmark of rejecting the estimate predicted by the regression. This means if our p-value for an estimate is greater than 5% (0.05), we cannot construct a 95% confidence interval that contains the true value. We then would have insufficient evidence for our alternative hypothesis and fail to reject the null hypothesis that the coefficients impact on log odds could be 0. We observed that 4 out of 12 of our parameters had a p-value larger than 0.05. These parameters include people who felt reasonably safe walking home, feeling_of_safety_walking 2, people who did not walk home alone, feeling_of_safety_walking 5, people who are 25 to 34 years old, R_age_group 25 to 34, and people who are 35-44 years old, R_age_group 35-44. The parameters with large p-values had coefficient estimates that were much smaller and closer to 0 than the other coefficient estimates with acceptable p-values, further supporting the null hypothesis. While we cannot reject the null hypothesis we also cannot accept it. These estimates can not be used with confidence until more data is collected for further analysis. For the parameters with p-values less than 0.05 we claim that there is strong evidence against the hypothesis that our estimate equals 0. Estimates with low p-values can be used in our model with confidence to predict.

From our results and model, we can begin to take inferences that can apply to the "real world". As discussed above, it appears that being younger and being female is predictive of likelihood of taking a self defense course. This may be applied to marketing for martial arts and other combat schools, as well as law enforcement groups. All of these groups may want to know who is the most likely to take their course, so that they may identify and contact these demographics to boost their income, profit, popularity or to forward their organizational goals, such as increasing feelings of safety in their neighbourhoods. Identifying demographics that may be more likely to take a self defense course could lead to other studies, possibly examining the explanatory variables of feelings of safety and what is predictive of that. Another study may explore other similar courses or programs that respondents may be interested in, such as a firearms course.

When we compare our model to other studies we see that the slope parameters agree with other studies. In a gallup poll conducted in 2012, they studied how safe men and women feel in many developed countries and conducted surveys in 143 countries [Nsubuga, 2020]. What they found is that "men and women often do not equally share the bolstered sense of security that these types of improvements bring" (ie. the increased security of living in a wealthy country). Relating this to our model we found women are more likely to take a self defense course, which is consistent with the gallup polls finding that women are more likely to take steps to increase their feelings of safety.

Weaknesses

One issue with the survey is that the data is all categorical and not continuous or discrete, this issue in particular makes it difficult to do analyses as it forces us to use many dummy variables in our regression. As you can see looking at our equation it has 11 slope parameters. This happens as all the variables were categorical and required dummy variables for their outcomes. For example, we wanted to examine the effect age has on people's likelihood of taking a self defense course. Instead of just having one continuous variable and plugging in the person's age into our model we only had access to a categorical variable called age group. This affected our results by segmenting the population into large groups rather than individual years, possibly leaving notable age relationships out. One notable age grouping is the 15-24 year old group, which includes high school age kids, college age young adults, and college graduate age young adults. These are three distinct and different age groups that are lumped together by virtue of the data.

Another issue with the data is that there was a number of NA values meaning that they did not answer the question. This forces us to exclude approximately 0.4% of observations because of lack of response on our observed variables. Although it is not a significant number of observations it could have minor consequences affecting our estimates.

A third issue is that the participation from each age group does not seem to match the Statistics Canada age group statistics. Looking at the data from Stats Canada, it resembles all the age groups from 15-24 to 45-54 from the GSS Survey data, but then starts to decrease and deviate by a considerable margin for age groups after that. This means that the groups of 55-64, 65-74, and 75 and older were likely over-represented in the survey compared to the younger groups that were under-represented. This could be due to the fact the survey was voluntary and mostly done over the phone. This could be a result of people in the older age brackets being more likely to complete a phone survey compared to younger people. This could affect the results as we assumed the study to be a simple random sample with an accurate representative sample of the population of Canada.

Continuing on from our last issue with the way the responses were collected, this may be an issue and affect the surveys ability to obtain an accurate representation for all of Canada. It may be that the survey is not a true simple random sample. Younger generations seem less likely to answer the phone to do a survey. The GSS website states that in 2014 they began conducting some interviews online but this still results in a population distribution different from all of Canada. Phone surveys may not result in a perfect simple random sample and we adjusted our model based on it being simple random. The resulting effect on the estimates we obtained is that our finite population correction may be different, which would likely change the variance of our coefficients.

Next Steps

Comparing our results to others we see that our results are similar to other studies on the same subject matter and that our parameters of regression make sense. In a study done by YouGov they found that 61% of women take steps to avoid sexual assault [Ballard, 2019]. The study looks at ways in which women protect themselves and shows that in most ways women are more likely than men to find ways to protect themselves. In our next steps, we may use our model structure to find results on the other "feelings of safety" variables in the GSS data set, such as feeling of safety at home and feeling of safety on public transportation. Additionally, a new GSS survey covering new detailed or related variables would allow for new models to explore feelings of safety and methods of increasing confidence in personal safety in greater detail and precision. If the new survey has new continuous variables, or changes some categorical variables to continuous, such as age group to age, then it may open up other model types to explore this data and topic. These attributes of a new survey could also eliminate some of the weaknesses in the data format as listed above, such as strictly identifying the sampling technique, and taking extra care to grab a more representative population.

References

Ballard, J. (2019, March 28). 61% of women regularly take steps to avoid being sexually assaulted. Retrieved October 20, 2020, from https://today.yougov.com/topics/lifestyle/articles-reports/2019/03/28/women-safety-sexual-assault-awareness (https://today.yougov.com/topics/lifestyle/articles-reports/2019/03/28/women-safety-sexual-assault-awareness)

Buteikis, A. (2018). Practical Econometrics and Data Science. Retrieved October 20, 2020, from http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-5-Multiple-collinearity.html (http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-5-Multiple-collinearity.html)

Caetano, Alexander (2020), Various lecture coding

David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.1.

Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. Journal of the American Statistical Association, 87(417), 178-183. doi:10.2307/2290467 (doi:10.2307/2290467)

Government of Canada, S. (2019, February 20). Retrieved October 19, 2020, from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm (https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm)

Government of Canada, S. (2015, November 27). Table 2.1-1Annual population estimates by age group and sex at July 1, provincial perspective - Canada. Retrieved October 19, 2020, from https://www150.statcan.gc.ca/n1/pub/91-215-x/2014000/t512-eng.htm (https://www150.statcan.gc.ca/n1/pub/91-215-x/2014000/t512-eng.htm)

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/ (https://socialsciences.mcmaster.ca/jfox/Books/Companion/)

Kassambara, & U, M. (2018, March 11). Logistic Regression Assumptions and Diagnostics in R. Retrieved October 20, 2020, from http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/ (http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/)

Nsubuga, S. (2020, January 13). Women Feel Less Safe Than Men in Many Developed Countries. Retrieved October 20, 2020, from https://news.gallup.com/poll/155402/women-feel-less-safe-men-developed-countries.aspx (https://news.gallup.com/poll/155402/women-feel-less-safe-men-developed-countries.aspx)

T. Lumley (2020) “survey: analysis of complex survey samples”. R package version 4.0.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.