Fatal Shootings in the United States by Police Officers

Group 13

By: Carol Dang, Joy Mohr, Maybellyn Yap

Drexel University

BSAN360-001: Programming for Data Analytics

Cassidy Buhler

March 15, 2022

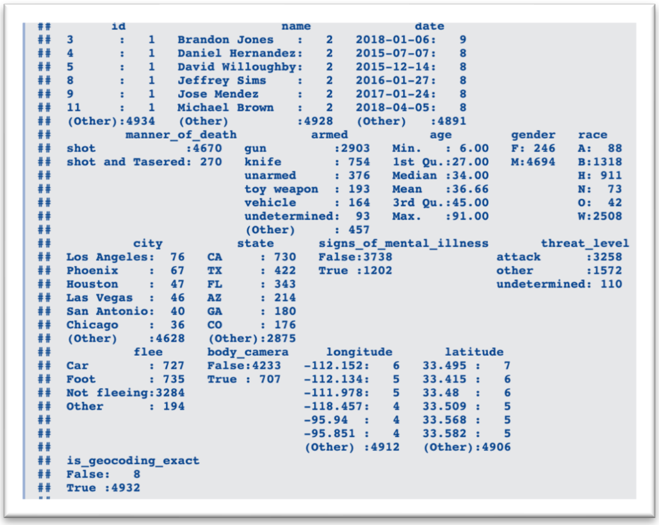
Introduction

We are aware of the rising fatal shootings by police officers that have sparked many historical protests. We witnessed how these fatal incidents have affected the lives of the victim's family members and the careers of the police officers. Our motivation is to assess and run analysis in hopes of gaining a better understanding of the underlying reasons that may have led to these unfortunate events.

For this project, our research questions are: **(1)** What is the average age of the victims for each gender? **(2)** Is the correlation coefficient between the age and mental illness of the victims statistically significant? **(3)** Is the correlation coefficient between the gender and mental illness of the victims statistically significant? **(4)** Are the differences between race with respect to their observed signs of mental illness statistically significant? **(5)** What is the average age of fatality in male compared to the average age of fatality in women? **(6)** Is personal profile (individual background) being a good predictor of a victim's mental illness. **(7)** Is personal profile (including health background and whether they are armed or unarmed during the incident) is a good predictor victim's manner of death (in this case, either shot or tasered and then shot). **(8)** Is a particular date and state (assuming state laws requiring police to carry a body camera to record the incident) is a good predictor of body camera.

Data

Our dataset introduces every record of fatal shootings in the United States by police officers who were in line of duty since Jan. 1, 2015. We obtained this data from the website www.github.com in which the owner of the database is The Washington Post. After omitting all our NAs, the dataset contains 17 variables, with a total of 4940 observations. The categorical variables in our data are: **(1)** id, **(2)** name, **(3)** date, **(4)** manner of death, **(5)** armed, **(6)** gender, **(7)** race, **(8)** city, **(9)** state, **(10)** signs of mental illness, **(11)** threat level, **(12)** flee, **(13)** body camera, **(14)** is geocoding exact, whilst our quantitative data are: **(1)** age, **(2)** longitude, **(3)** latitude. Since we have limited quantitative data, we then further converted some of our categorical variables into binaries to better run our analysis. Those converted variables are **(1)** manner of death, **(2)** gender, **(3)** signs of mental illness, **(4)** body camera. Our dataset summary is represented in Figure 1.



```
##         id                     name               date     
##  3      :   1   Brandon Jones   :   2   2018-01-06:   9  
##  4      :   1   Daniel Hernandez:   2   2015-07-07:   8  
##  5      :   1   David Willoughby:   2   2015-12-14:   8  
##  8      :   1   Jeffrey Sims    :   2   2016-01-27:   8  
##  9      :   1   Jose Mendez     :   2   2017-01-24:   8  
##  11     :   1   Michael Brown   :   2   2018-04-05:   8  
##  (Other):4934   (Other)         :4928   (Other)   :4891  
##           manner_of_death          armed            age          gender   race    
##  shot            :4670   gun         :2903   Min.   : 6.00   F: 246   A:  88  
##  shot and Tasered: 270   knife       : 754   1st Qu.:27.00   M:4694   B:1318  
##                          unarmed     : 376   Median :34.00            H: 911  
##                          toy weapon  : 193   Mean   :36.66            N:  73  
##                          vehicle     : 164   3rd Qu.:45.00            O:  42  
##                          undetermined:  93   Max.   :91.00            W:2508  
##                          (Other)     : 457                                    
##           city           state      signs_of_mental_illness       threat_level 
##  Los Angeles:  76   CA     : 730   False:3738                attack      :3258  
##  Phoenix    :  67   TX     : 422   True :1202                other       :1572  
##  Houston    :  47   FL     : 343                             undetermined: 110  
##  Las Vegas  :  46   AZ     : 214                                                
##  San Antonio:  40   GA     : 180                                                
##  Chicago    :  36   CO     : 176                                                
##  (Other)    :4628   (Other):2875                                                
##           flee       body_camera     longitude        latitude    
##  Car        : 727   False:4233   -112.152:   6   33.495 :   7  
##  Foot       : 735   True : 707   -112.134:   5   33.415 :   6  
##  Not fleeing:3284                -111.978:   5   33.48  :   6  
##  Other      : 194                -118.457:   4   33.509 :   5  
##                                  -95.94  :   4   33.568 :   5  
##                                  -95.851 :   4   33.582 :   5  
##                                  (Other) :4912   (Other):4906  
##  is_geocoding_exact
##  False:   8        
##  True :4932        
```

Figure 1

Analyses and Results

Analysis 1: What is the average age of the victims for each gender?

By running the code aggregate(formula=age~new_gender, FUN=mean, we retrieved the result of the average age of victim for male is 37 while female is 36. Do note that the converted variable new_gender is from its original variable gender. Figure 2 shows the overall age demographic of victims by gender.

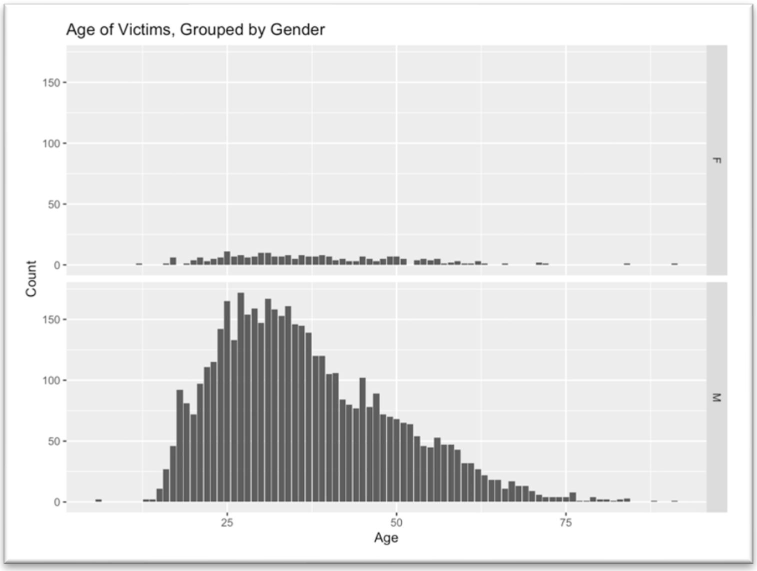


Figure 2

Analysis 2: Is the correlation coefficient between the age and mental illness of the victims statistically significant?

After running a correlation test using the code cor.test(age,new_signs_of_mental_illness), we retrieved the result of p-value = 1.143e-13, which is less than 0.05, therefore rejecting our null hypothesis of there is no difference between the age and mental illness of the victims in a statistically significant way.

Analysis 3: Is the correlation coefficient between the gender and mental illness of the victims statistically significant?

After running a correlation test using the code cor.test(new_gender,new_signs_of_mental_illness), we retrieved the result of p-value = 0.0002316, which is less than 0.05, therefore rejecting our null hypothesis of there is no difference between gender and mental illness of the victims in a statistically significant way.

Analysis (4): Are the differences between race with respect to their observed signs of mental illness statistically significant?

To answer and observe this question, a bar graph was first created to observe if there were any possible visual indications of a relationship between the variables race and mental illness = (True). There were observed differences between races and mental illness = (True) so a Chi-Square Test and was used to determine if there was relationship. A Chi-Square test was then chosen for this analysis because both variables being tested were discrete. The result of the Chi-Square test concluded a p-value of 2.2e-16, which is < 0.05. With 95% confidence, we can reject the null hypothesis and conclude that the victim's race and their sign of mental illness = (T) are statistically significant from each other. Do note that in this analysis, we converted signs of mental illness to binaries, with True = 1 and False = 0. Figure 3 describes the mental illness grouped by race.

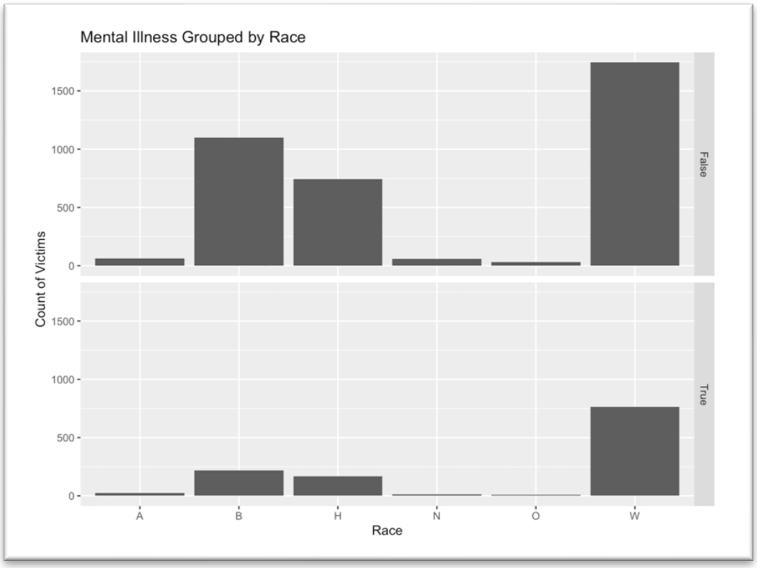


Figure 3

Analysis (5): Is the average age of fatality in gender (M) significantly similar than the average age of fatality in women?

All necessary variables were transformed to ensure that they were normally distributed. This was done because T-tests assume that the mean of the sample is normally distributed, if the variables were not transformed then the results would be inaccurate. An ANOVA Test and T-test were chosen for this analysis. A T-test was appropriate for this test due to its ability to compare two segments against each other. A T-test is also used for continuous data which Age was a variable being tested. A one-way ANOVA test was also used to assess whether the groups being compared differ from one another based on one factor (age). The result of the t-test concluded a p-value of 0.1456, which is > 0.05. The result of the ANOVA test concluded a p-value of 0.1374, which is > 0.05. With 95% confidence, we fail to reject the null hypothesis and conclude that the average age of fatality in men is statistically similar to the average age in fatality in women.

Analysis (6): Is personal profile (individual background) being a good predictor of a victim's mental illness?

We ran a multivariate linear regression model to observe whether an individual background factors such as age, gender, race, and city could predict the victim's mental illness. We obtained the result of adjusted R-squared of 0.1067, indicating the model explains 10.67% of the variation in the data.

Analysis (7): Is personal profile (including health background and whether they are armed or unarmed during the incident) is a good predictor victim's manner of death (in this case, either shot or tasered and then shot)?

We incorporate analysis 6 with additional factors such as signs of mental illness and whether they were armed at the time of incident to determine if all factors combined could predict the victim's manner of death. There are only 2 outcomes for manner of death, which are shot and tasered then shot. We obtained the result of adjusted R-squared of 0.0974, indicating the model explains 9.74% of the variation in the data.

Analysis (8): Is a particular date and state a good predictor of body camera?

Assuming state laws requiring police to carry a body camera to record the incident, we used factors such as date and state to determine if those factors could predict whether the police officer is carrying a camera at the time of incident. We obtained the results of adjusted R-squared of 0.0587, indicating the model explains 5.87% of the variation of the data.

Conclusion

After running multiple methods of analysis, we observed the correlation coefficient for age and mental illness, as well as gender and mental illness are statistically significant. From the result of p-value less than 0.05, we conclude that there is a linear relationship between age and mental illness, as well as gender and mental illness. Furthermore, the victim's race and their signs of mental illness (True or 1), are statistically significant from each other. We also conclude that the average age for fatal shootings of men is significantly similar to the average age for fatal shootings of women. From a wider view of the multivariate linear regression models, the models did not explain the variation in the data well enough as their adjusted R-squared value is too low, averaging at less than 9%. However, by comparing the models, model 1 (analysis 1) ranked first, followed by model 2 (analysis 2), and lastly model 3 (analysis 3). The difference between the variation of these models is very low as well.

Because of the nature of our data and our limited knowledge in coding, we face limitations in running analysis because of the lack of numerical data. However, we would like to highlight that converting some of our variables to binaries has helped tremendously in conducting our analysis. We learned that the insights that are provided throughout this report may be used to help educate the public on these fatal incidents and inform legislature or authoritative bodies on the significant findings to help facilitate outreach programs to limit or decrease the amount of these fatalities. In conclusion, we discern that data ethics will be a huge part of our career, and we appreciate that we are given the opportunity to work on this data and this experience will remain to remind us to uphold our integrity and carry out research without any bias down the road.

Bibliography

WP Company. (2020, January 22). *Fatal force: Police shootings database*. The Washington Post. Retrieved March 16, 2022, from https://www.washingtonpost.com/graphics/investigations/police-shootings-database/