The Macabre Reality of Gun Deaths in the US by Aditya Kedia

Introduction

As a foreigner in the US, the relationship of the American diaspora with firearms, often bordering on grim obsession, has always surprised me. Private gun ownership in the US is the highest of any country in the world, and the fact that this would lead to more firearm deaths is rather obvious. Often, however, the focus of these discussions is mainly concentrated on the perpetrators of gun violence, and not enough attention is paid to the deaths and the victims themselves. In this document, therefore, I aim to explore these deaths form the years 2012-2016, and find out more about the victims themselves.

Acquiring The Data

The data was acquired from the CDC’s Multiple Cause of Death datafile. It was parsed using a script obtained from the wonderful data journalism site fiveThirtyEight.com. The Script, and the CSV file containing the data for all the years collected in one csv file are included with this submission.

Univariate Plots Section

Let us dive into some summaries of and inital looks at the data.

```
## 'data.frame':    175708 obs. of  11 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ year     : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ month    : int  1 1 1 2 2 2 2 3 2 2 ...
##  $ intent   : Factor w/ 4 levels "Accidental","Homicide",..: 3 3 3 3 3 3 4 3 1
3 ...
##  $ police   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ sex      : Factor w/ 2 levels "F","M": 2 1 2 2 2 2 2 2 2 2 ...
##  $ age      : int  34 21 60 64 31 17 48 41 50 NA ...
##  $ race     : Factor w/ 5 levels "Asian/Pacific Islander",..: 1 5 5 5 5 4 5 4 5
2 ...
##  $ hispanic : int  100 100 100 100 100 100 100 100 100 998 ...
##  $ place    : Factor w/ 10 levels "Farm","Home",..: 2 9 4 2 4 2 2 2 4 2 ...
##  $ education: Factor w/ 4 levels "BA+","HS/GED",..: 1 4 1 1 2 3 2 2 4 NA ...
```

The set contains 175708 gun deaths from 2012 to 2016. That’s 35141 gun deaths per year. Each entry contains the year and the month of the death, along with the age, sex, race and education level of the victim. If the victim was of Hispanic descent, the data gives their country of origin via a these codes. Additionally, it contains the place of death and whether the death was police related.

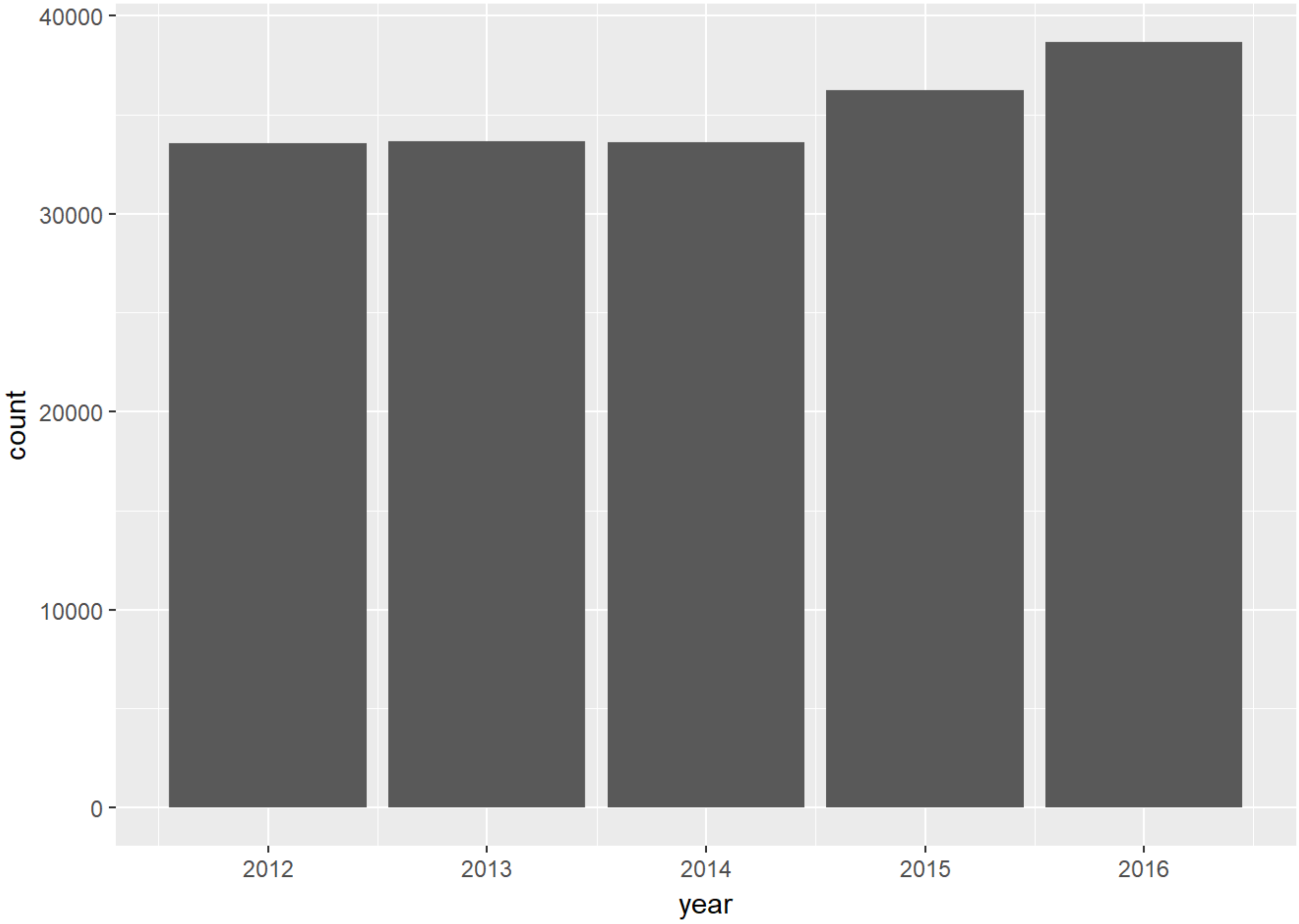
The index number, ‘X’ is a redundant variable and I will drop it in the next line of code. I am going to change the police variable to a factor rather than an integer. I’ll do the same for the ‘Hispanic’ variable and then look at a summary of the data

```
##       year          month                 intent       police
##  Min.   :2012   Min.   : 1.000   Accidental  :  2623   0:173312
##  1st Qu.:2013   1st Qu.: 4.000   Homicide    : 63559   1:  2396
##  Median :2014   Median : 7.000   Suicide     :108131
##  Mean   :2014   Mean   : 6.574   Undetermined:  1389
##  3rd Qu.:2015   3rd Qu.: 9.000   NA's        :     6
##  Max.   :2016   Max.   :12.000
##
##  sex             age                                     race
##  F: 25333   Min.   :  0.00   Asian/Pacific Islander        :  2420
##  M:150375   1st Qu.: 27.00   Black                         : 42234
##             Median : 41.00   Hispanic                      : 16125
##             Mean   : 43.58   Native American/Native Alaskan:  1662
##             3rd Qu.: 58.00   White                         :113267
##             Max.   :107.00
##             NA's   :26
##     hispanic                     place              education
##  100    :159058   Home               :91215   BA+         :22274
##  210    :  9982   Other unspecified :36388   HS/GED      :74927
##  260    :  1545   Other specified   :21183   Less than HS:38041
##  282    :   757   Street             :16836   Some college:37914
##  270    :   703   Trade/service area: 5149   NA's        : 2552
##  281    :   582   (Other)            : 2589
##  (Other):  3081   NA's               : 2348
```

Already we can discern the following facts -

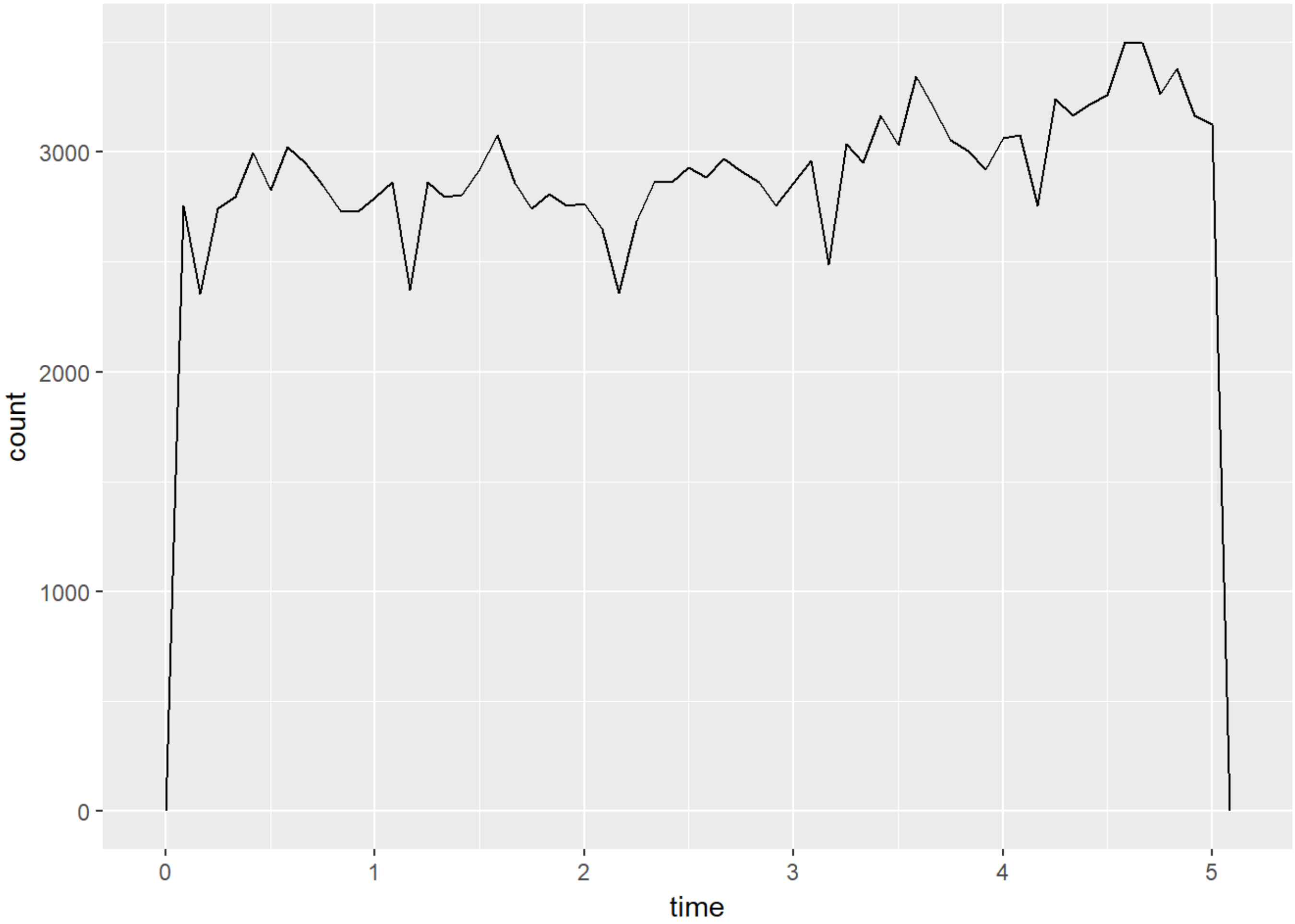
- Suicides far surpass accidents and homicides
- A vast majority of gun deaths are not police related
- More men die of bullet wounds than women
- People of Caucasian descent are overrepresented compared to other races
- Amongst Hispanics, Mexicans and Puerto Ricans are over-represented. This is likely as they are probably over-represented in the general population as well
- Majority of Gun deaths occur at home
- Most victims have a Highschool/GED education only

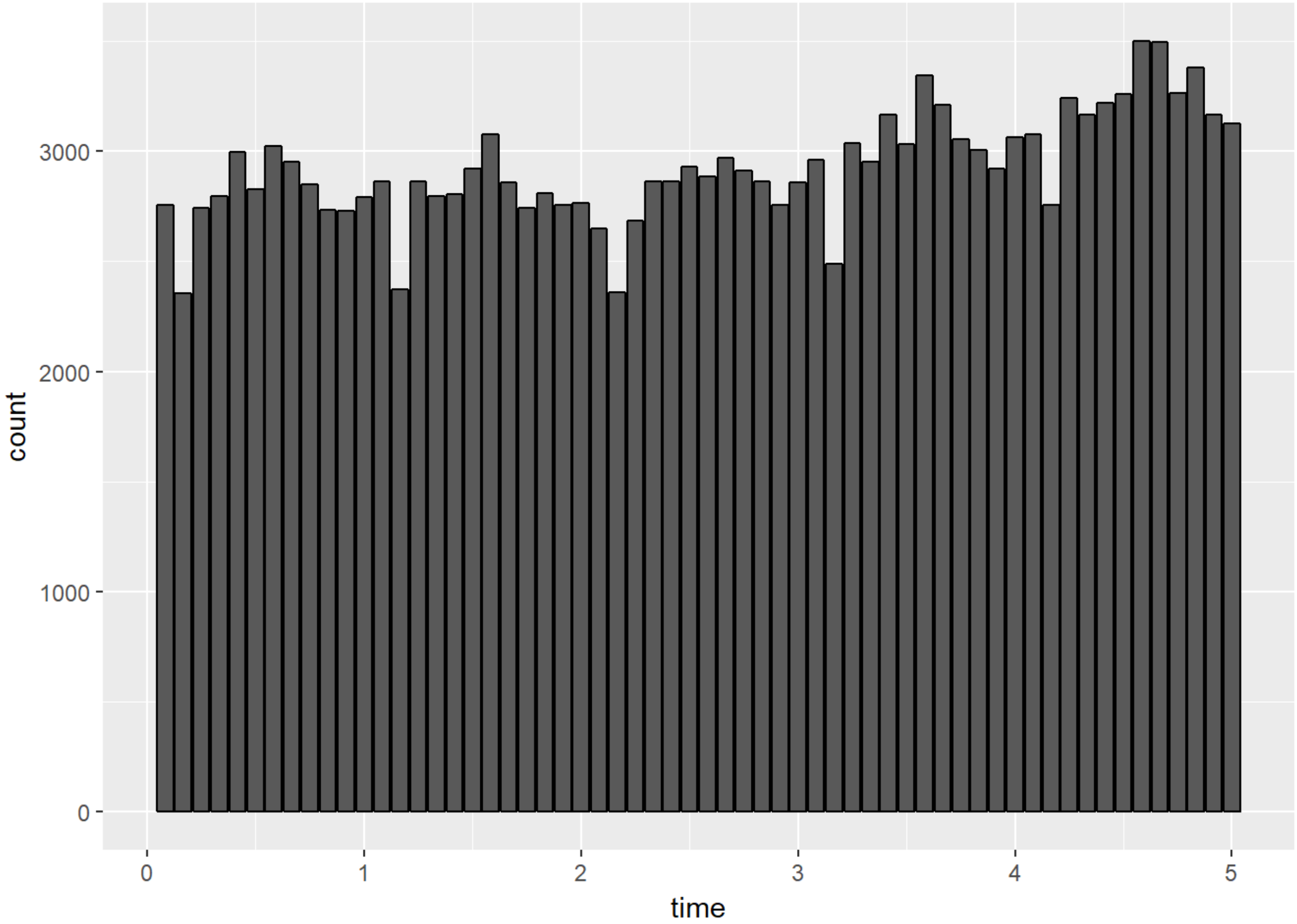
Next we will look at a bar plot of the data year by year.



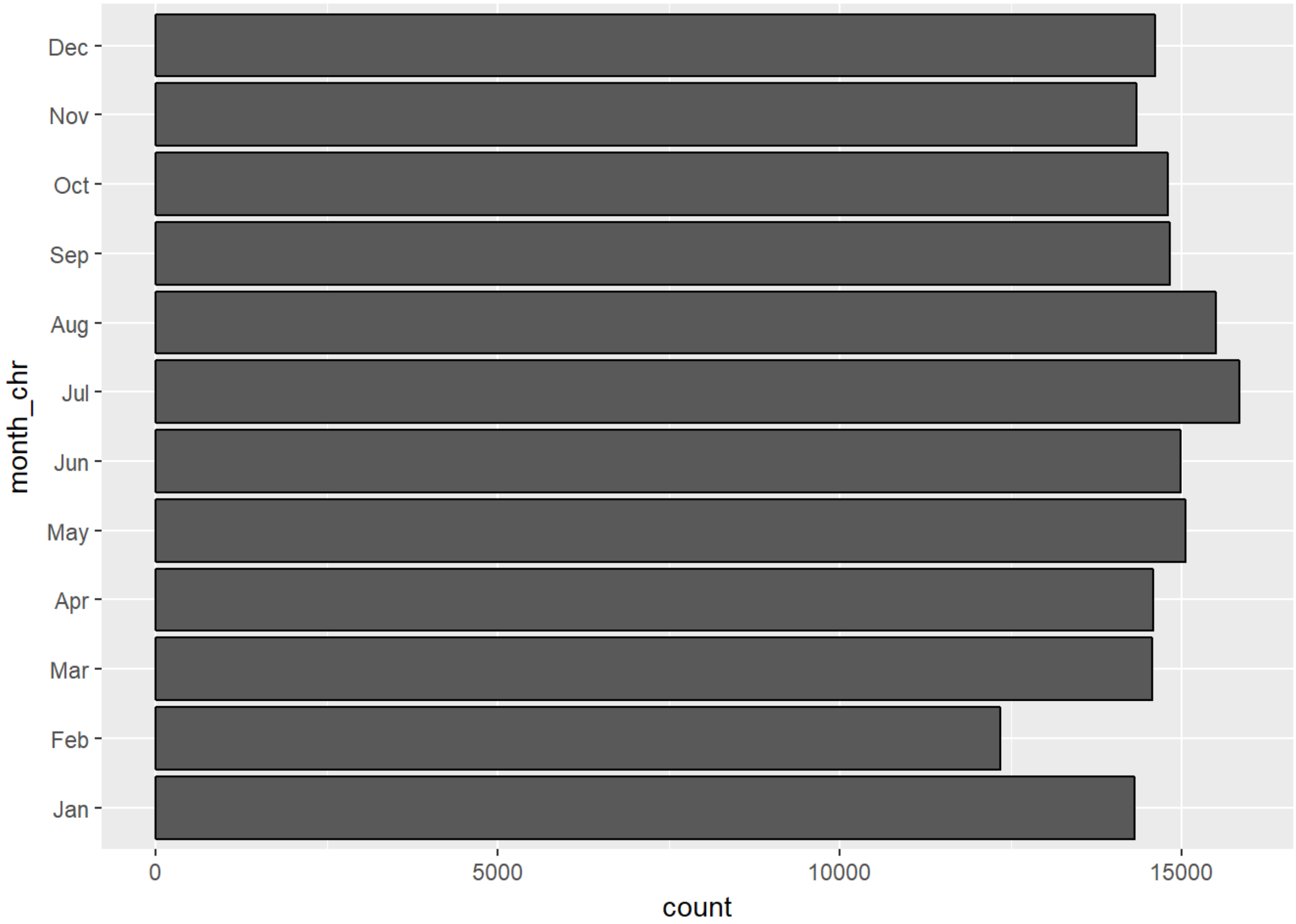
Gun deaths in the US stayed almost constant from 2012 to 2014, followed by successive increases in 2015 and 2016. Of course, these numbers would make more sense in the context of the actual population number. We will explore this further in the next section.

I am going to create a new variable using years and months such that “time” becomes a little more continuous. Using 2012 as the base year, and dividing the months by 12.



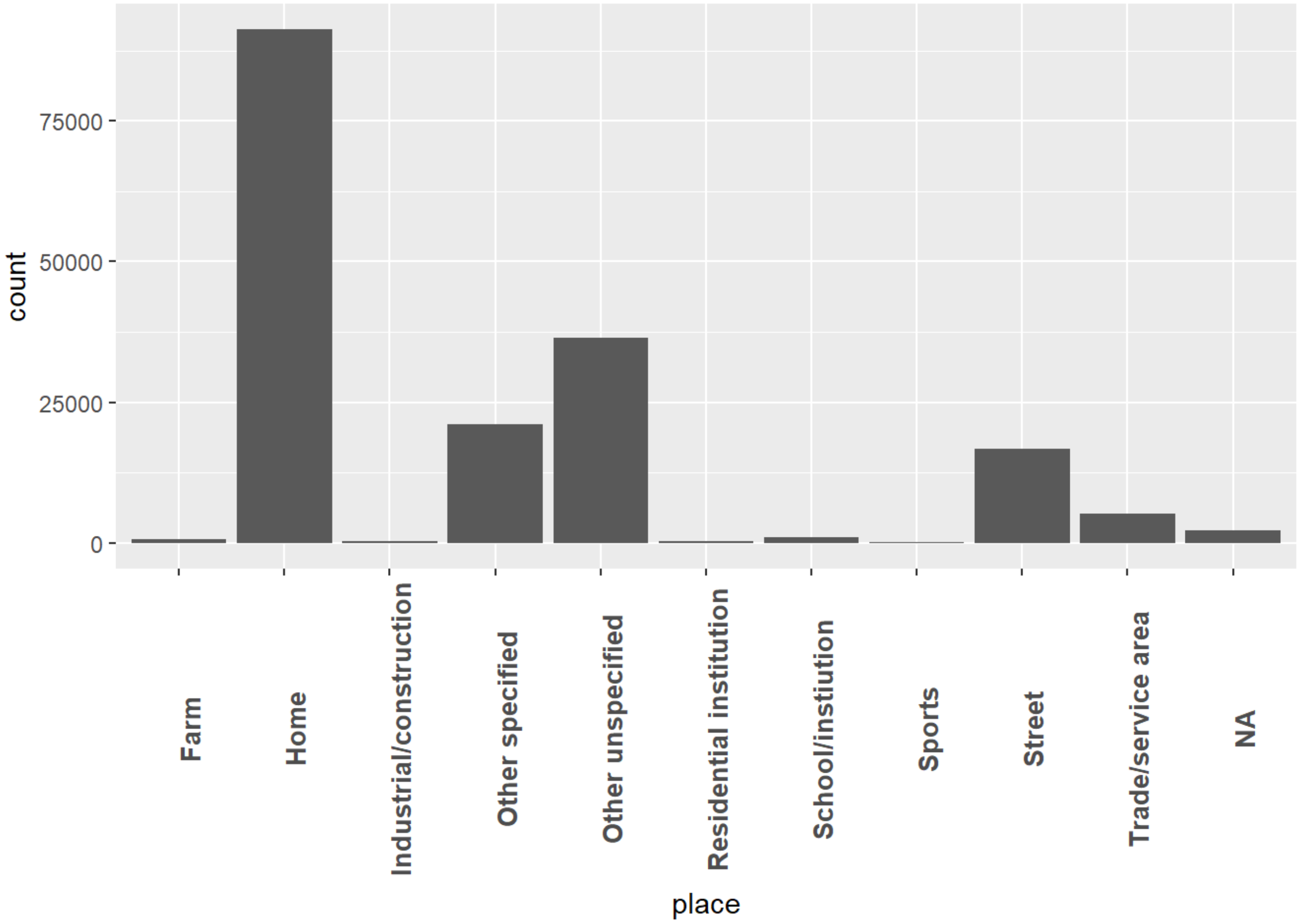


There seems to be a consistent dip in the earlier part of every year. I wonder if this pattern will persist if we plotted the count over the months

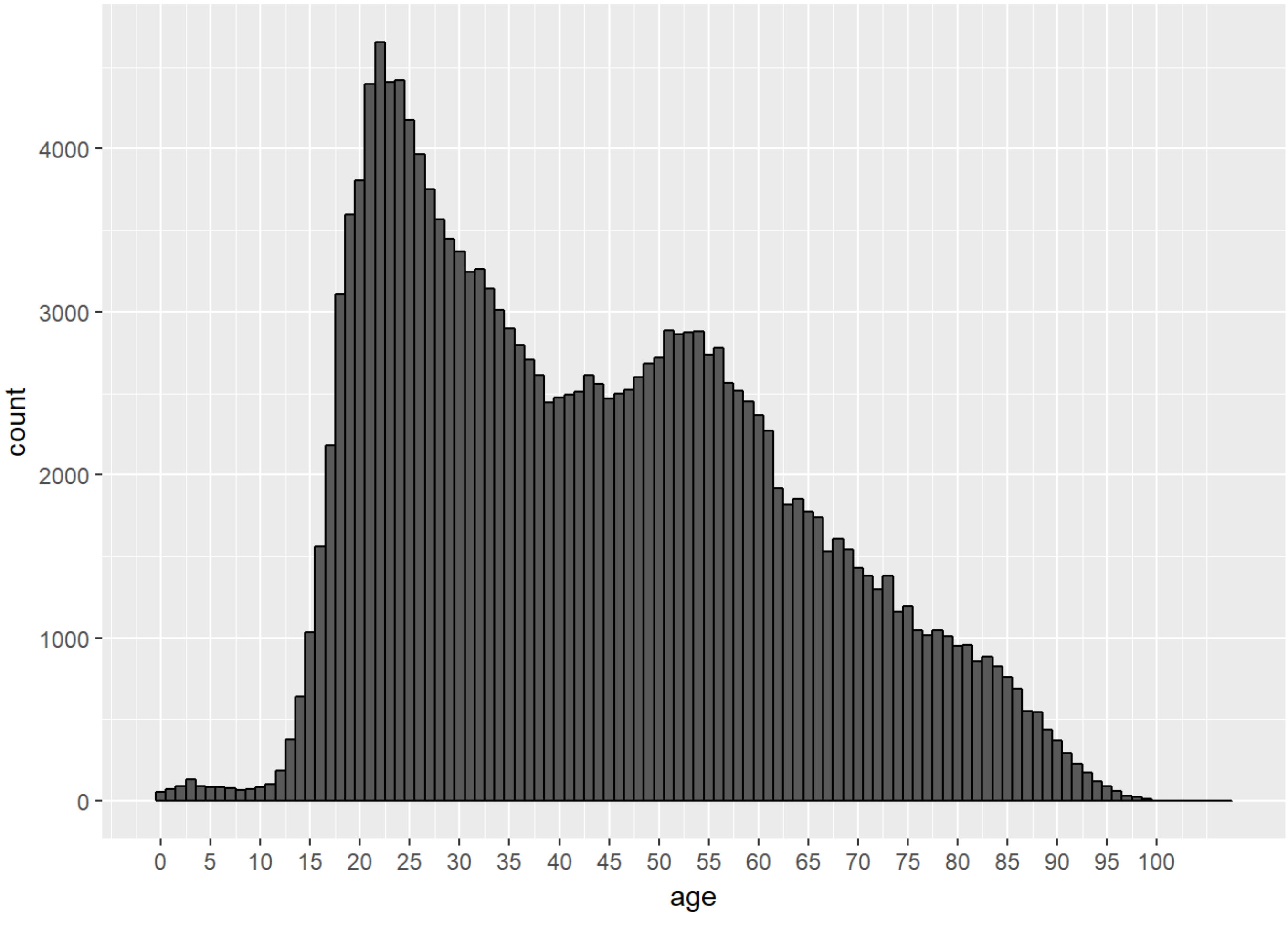


If there is a pattern to be gleaned here, it seems that there are consistently fewer deaths in February. This may be due to the fact that February has fewer days (28/29 as opposed to 30/31). Since the data across days is unavailable, it is hard to tell if there is a February effect.

Teh place of death also seems like an importnat variable, and the next plot compares the several factors using aother bar chart.



Let us also quickly look at distribution across age before we move on to bivariate analysis.



Looks like most victims of gun violence expire in their early 20s, and there is a sudden spike in the mid to late 50s. I am curious to see if there is a distribution across ages of the different intents.

I retrieved population numbers for the US from the census website so that we can look at deaths as a proportion of the population.

```
##   year       pop
## 1 2012 313993272
## 2 2013 316234505
## 3 2014 318622525
## 4 2015 321039839
## 5 2016 323405935
```

Univariate Analysis

What is the structure of your dataset?

The dataset contains several categorical variables for each death gun death in the US over the last 5 years (2012-2016). it classifies these deaths by intent, age, race, sex, place of death and education. additionally, it provides information about whether the death was caused by a police officer.

What is/are the main feature(s) of interest in your dataset?

I think the main feature of interest is the count itself and how it varies over the several categories

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

All the categorical variables will shed light on the count of gun deaths, why they happen, when, and to whom.

Did you create any new variables from existing variables in the dataset?

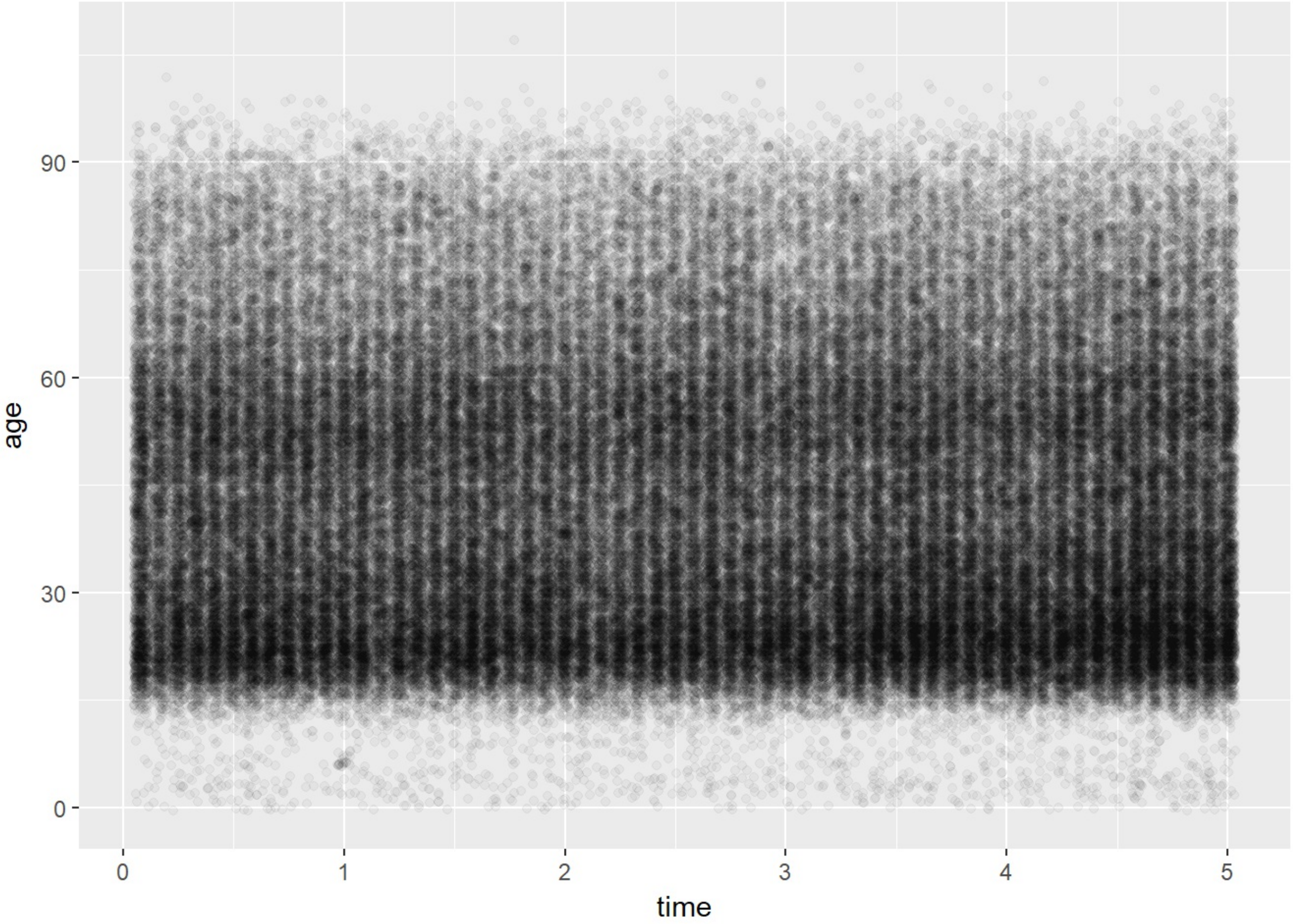
The dataset is severely lacking in continuous numerical variables. I, therefore, one called time, which is a combination of month and year. The more accurate way of looking at the count is to look at it as a function of the population of the country. I therefore acquired the population data, to be later added to grouped versions of the gun_death data frames.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

The dip in the number of deaths at the beginning of each year was a little surprising, although maybe this is due to the smaller number of days in February. The data for age vs gun deaths is nearly bimodal. I did not need to adjust or change the data yet, although I will need to group and summarize in later sections.

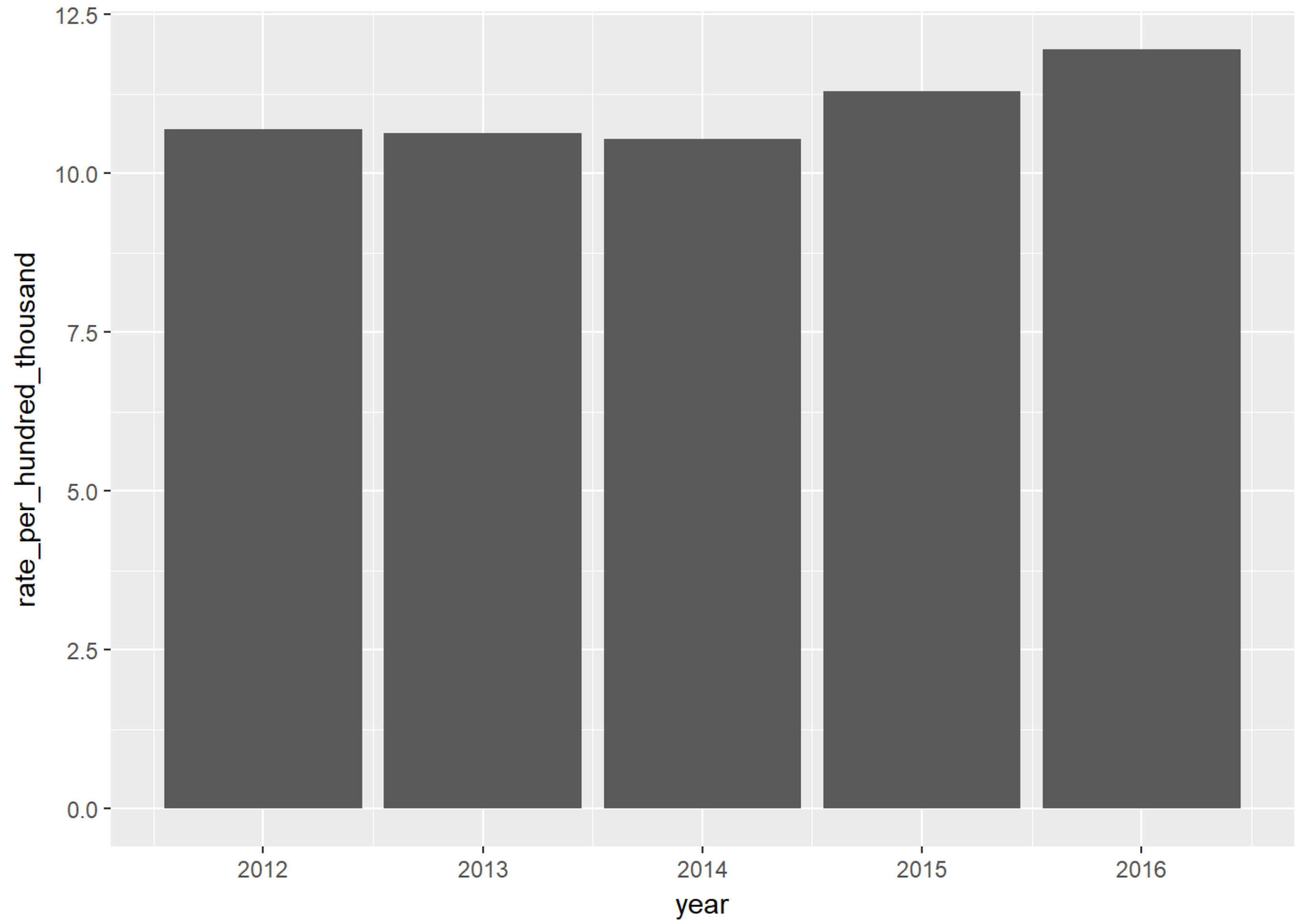
Bivariate Plots Section

The first bivariate plot I want to look at is that of time vs age, to get an idea of whether the age of the gun victim has changed at all over the 5 years.

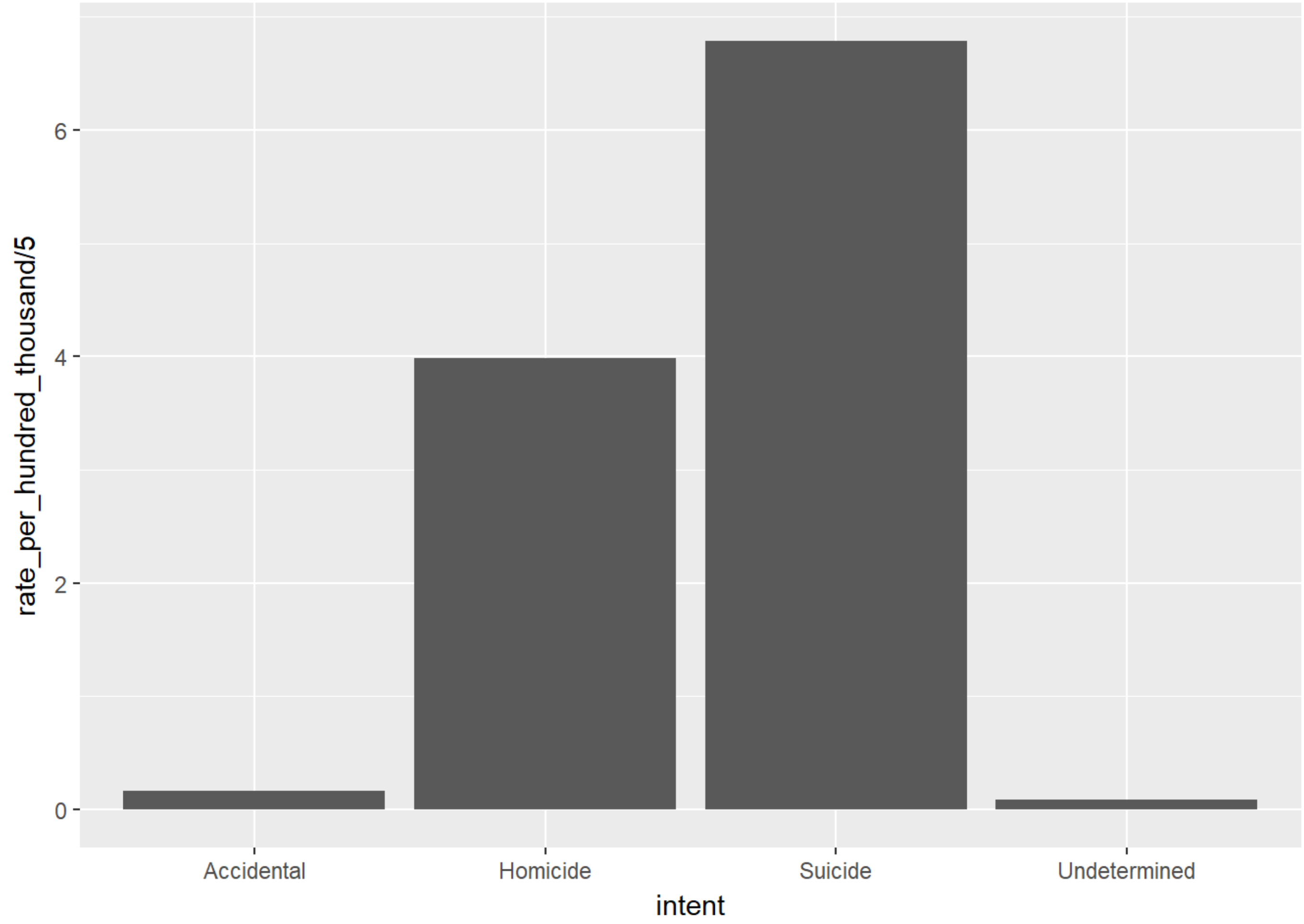


Note that in the above plot, the x-axis denotes time over 5 years, so 0 is the beginnig of 2012, while 5 is the end of 2016. No obvious patterns emerge here, suggesting that the distribution of ages of the victims of gun violence didn't change over the years. The two dark bands in the 20s and the 50s reinforces the earlier observation of the peaks in the count vs age distribution

Let us now group data by the years, create a new variable to see the change in rate of gun violence

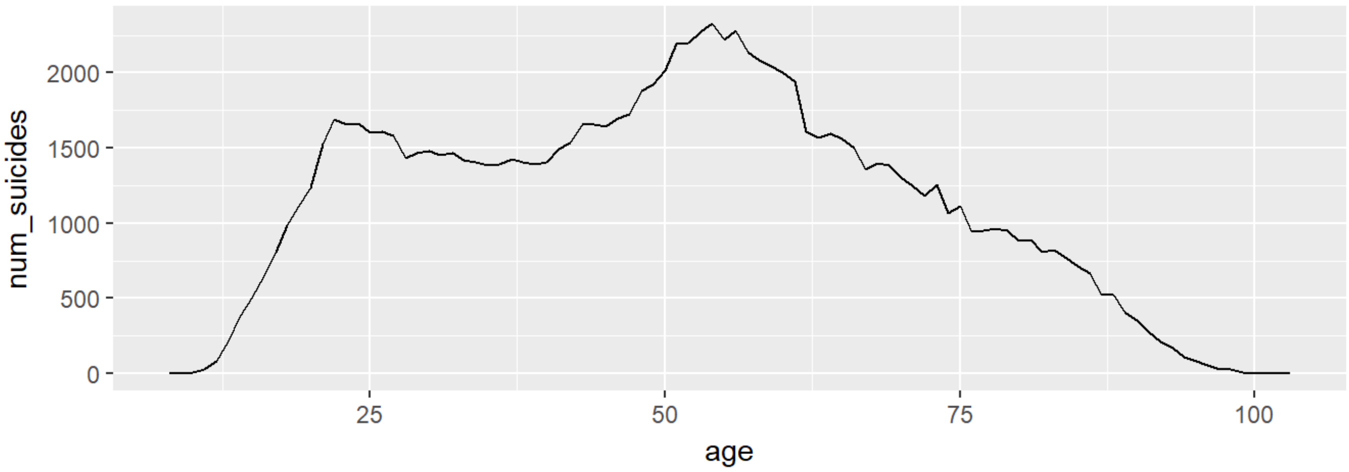


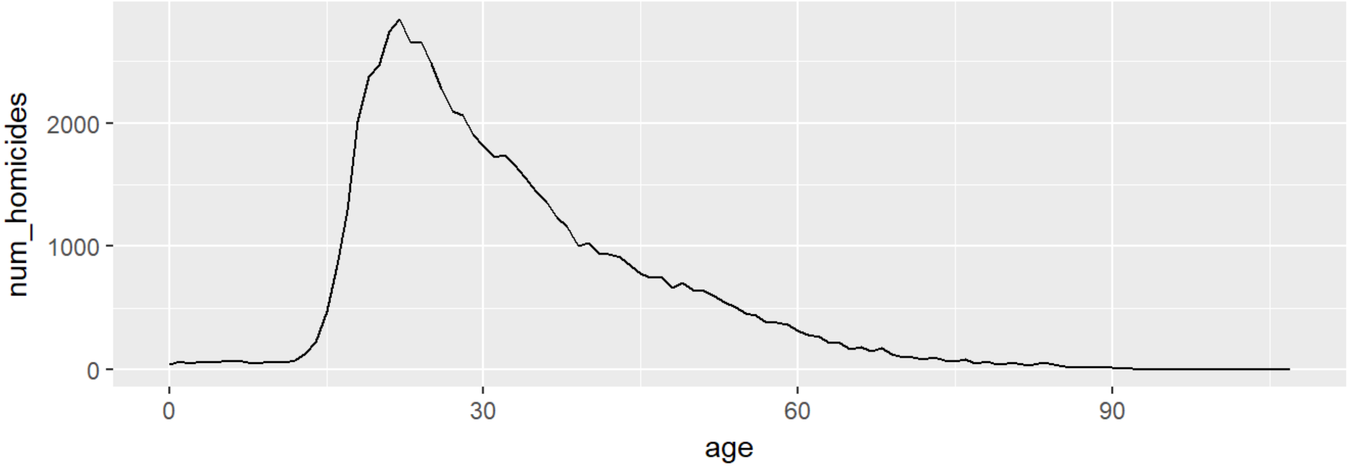
Now we can truly see that the death rate due to firearms has in fact increased in 2015, and then again in 2016. a similar chart should be constructed for the reason of death.



Again, we see that the number of deaths due to suicide is much greater than the number for any other reason.

Next, let us look at the correlation between age and suicide. In the next block of code, I have grouped the dataset by age and counted the number of suicides for each.





The above plots, while morbid and depressing, seem to reveal crucial truths about our society. earlier we saw how the distribution of gun deaths across the ages was distributed such that younger victims were over-represented in the sample. We now see that a majority of that number comes from homicides. Take them away and we see a very different pattern. People in their 50s are way more likely to commit suicide, while the younger victims of firearms are much more likely to die in a homicide.

I wonder if this pattern changes with race. We will explore this in the multi variate section.

We have other variables to analyze, and I think they would be better analyzed in a multivariate fashion.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

The number of gun deaths has increased in the last few years, and the trajectory seems upward. Suicides are much more common amongst firearm deaths than homicides. When looking only at the death due to suicides, people in their 50s are over-represented, while people in their 20s are over-represented in the population of younger victims.

Did you observe any interesting relationships between the other features
(not the main feature(s) of interest)?

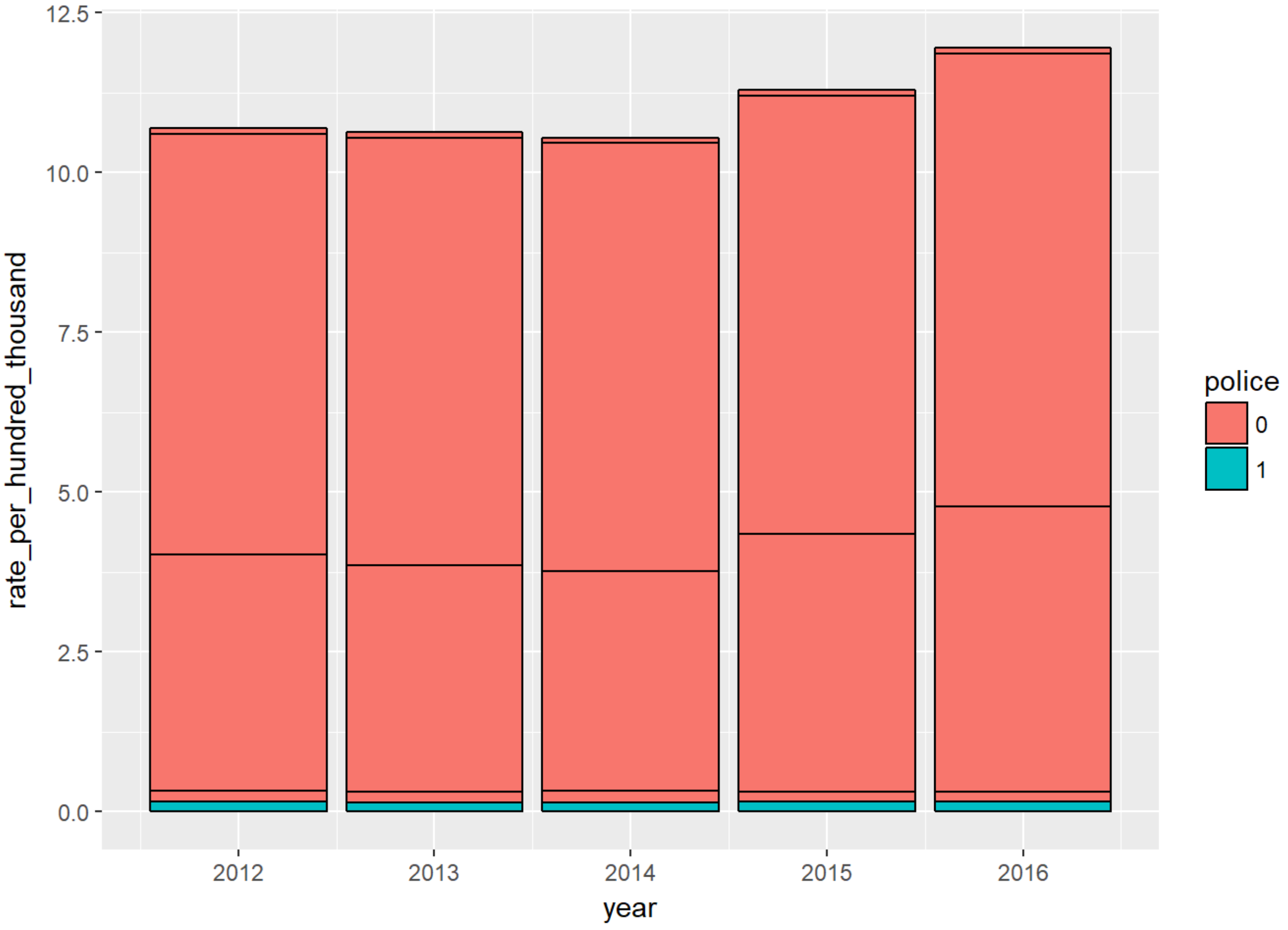
No particularly interesting relationships were observed amongst the other variables.

What was the strongest relationship you found?

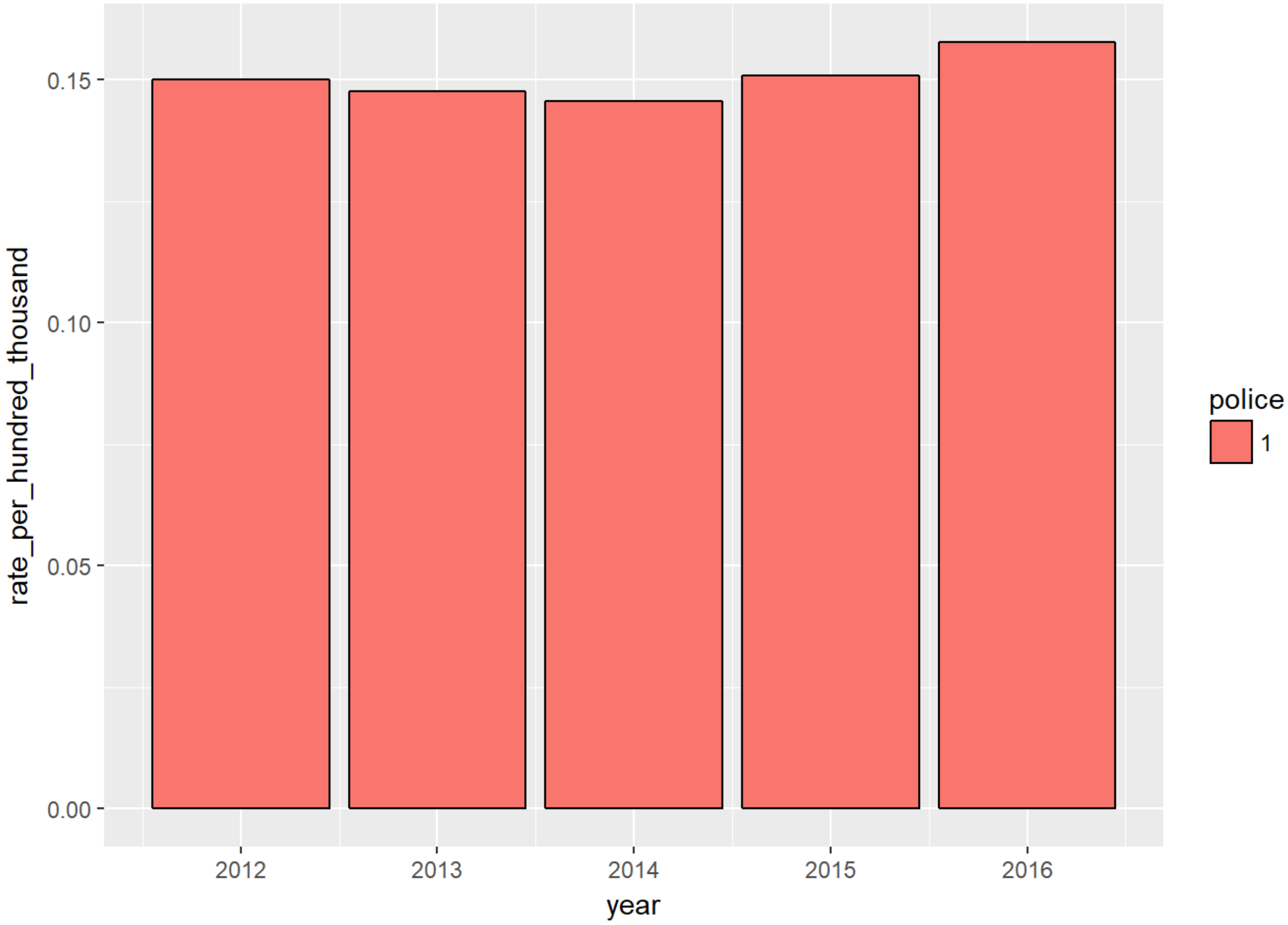
The strongest relationship is that amongst age and gun suicides, in that middle age commit gun suicides more often than any other age group.

Multivariate Plots Section

The first thing to explore here is the rate of gun death per year, colored by the the police variable, to see if ther have been changes in police gun deaths over the years.

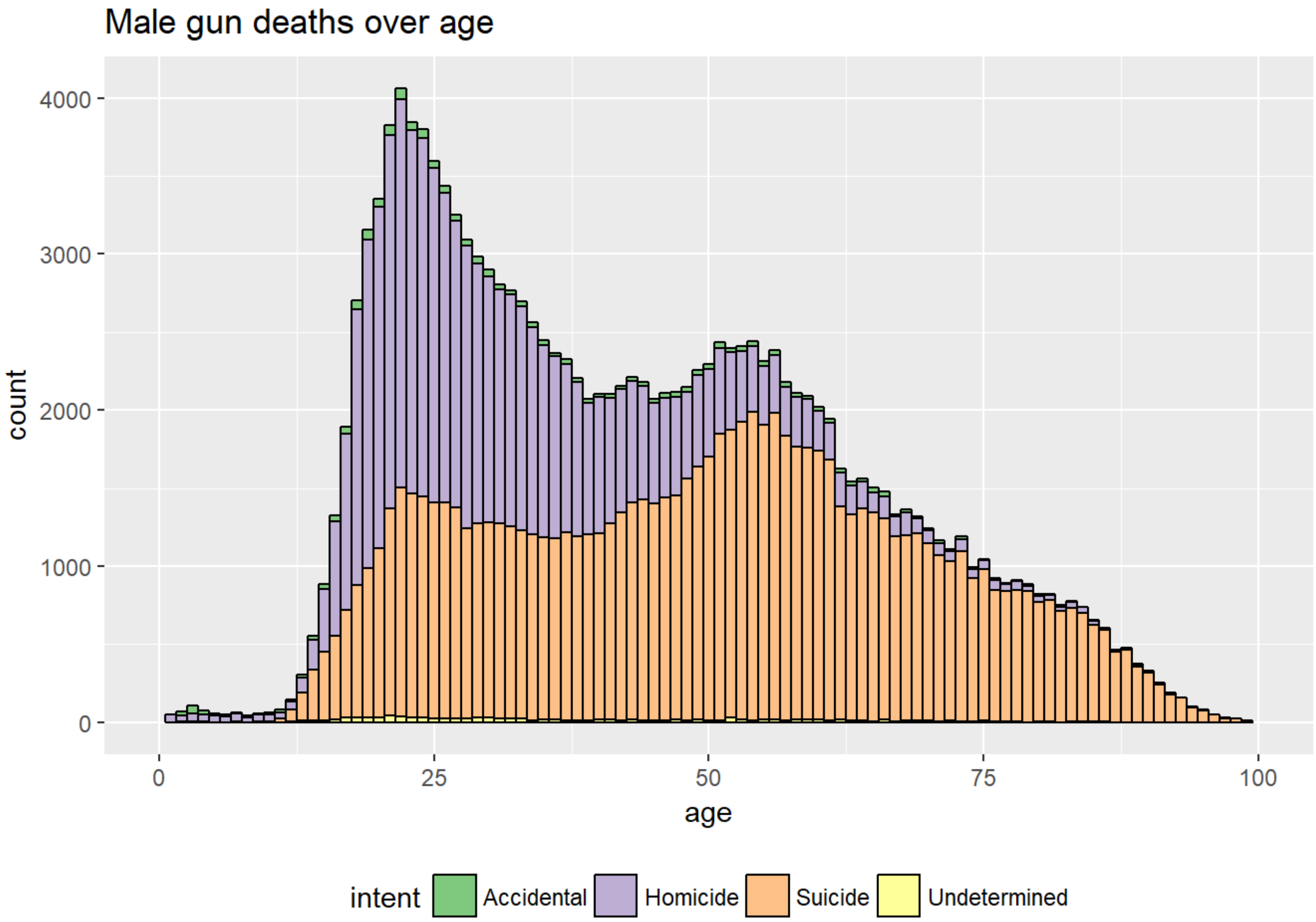


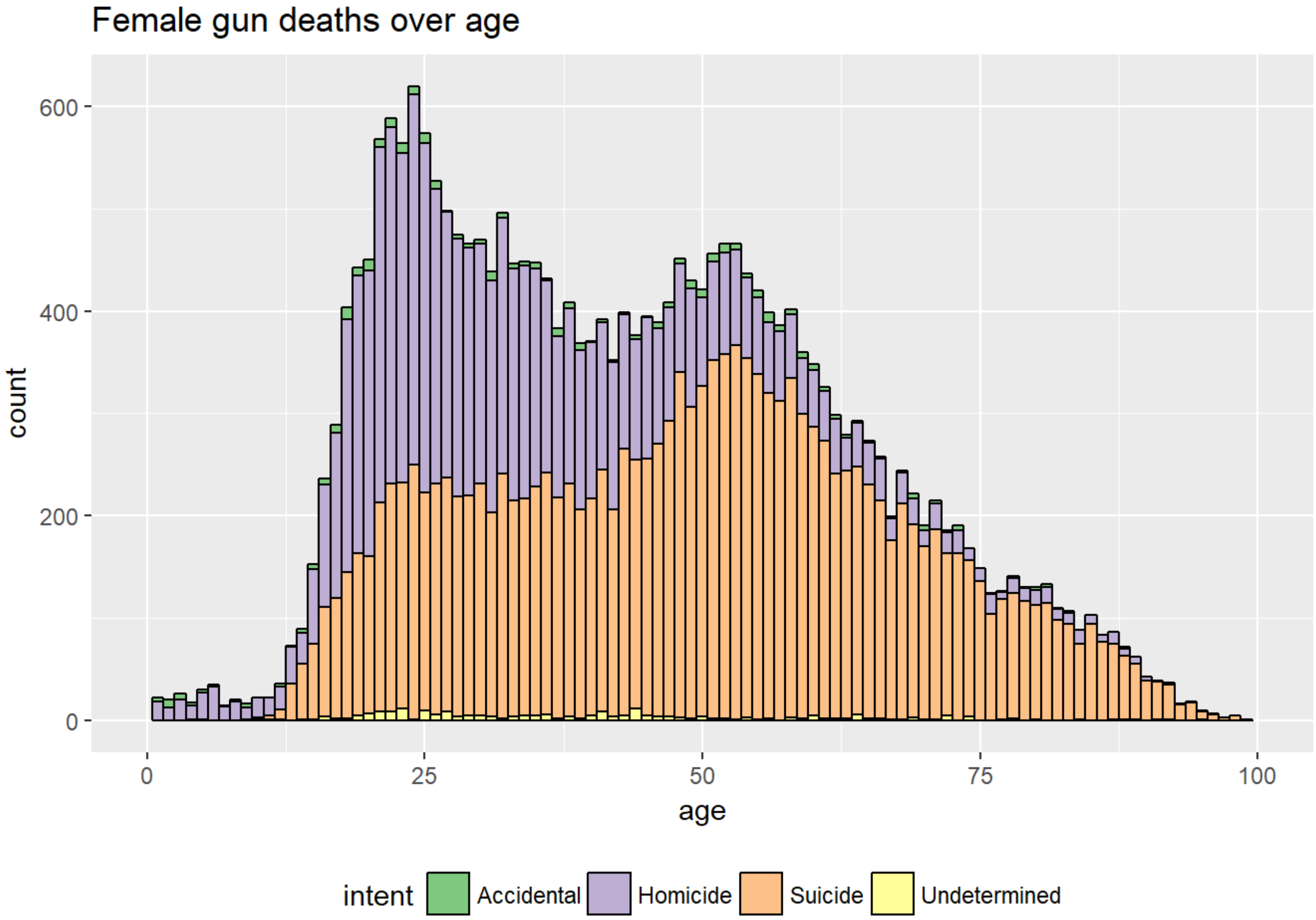
Looks like the number of non police deaths is so uch higher than the number of police deaths, that any pattern in the police deaths is drwoned out. Let us subset the data to see only the deaths caused by police, and try again.



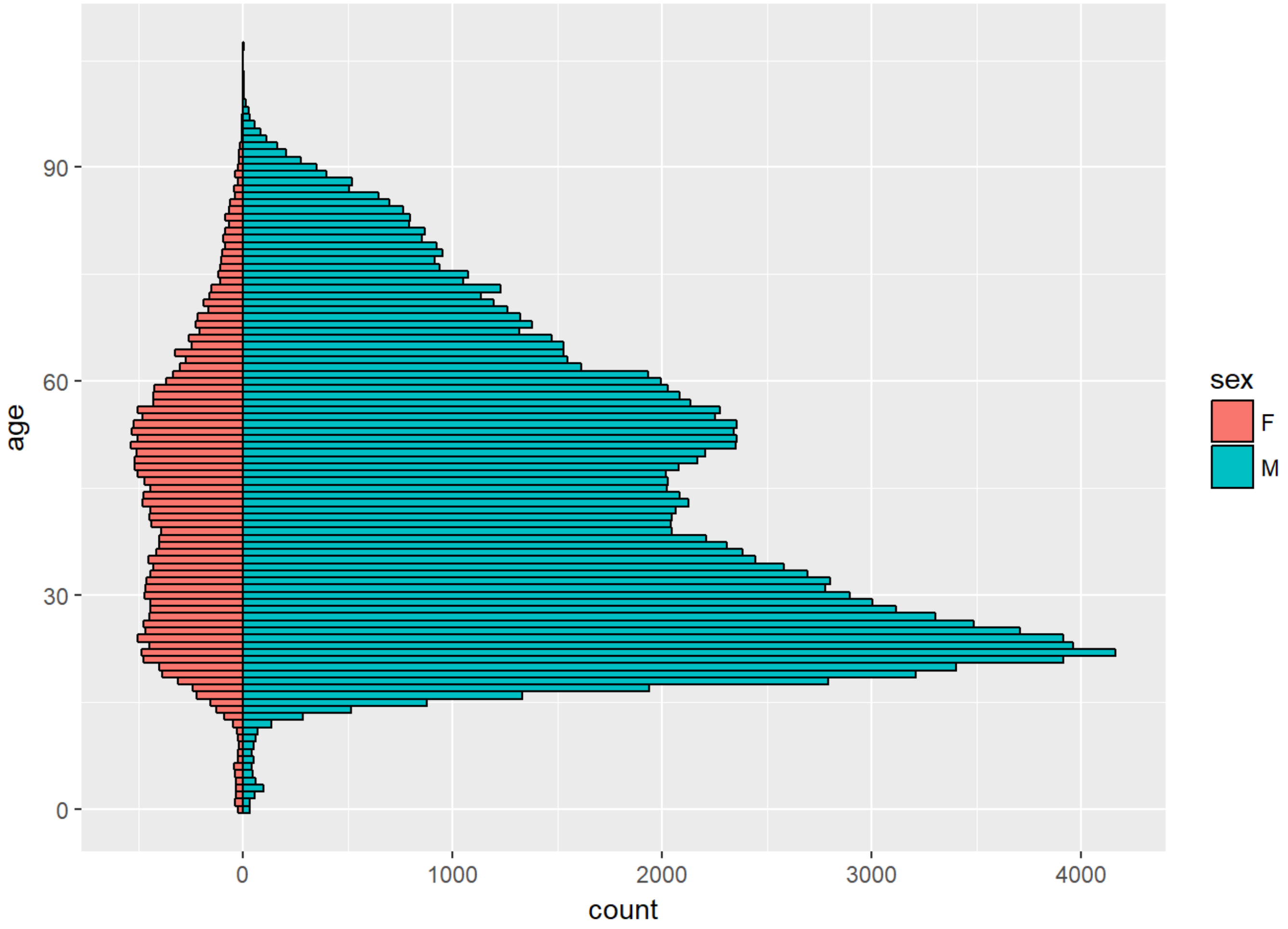
Now we get a better picture of the police deaths over the years. While there are some changes in the numbers over the years, they seen to small to actually find a pattern/trend in them.

Now we will flesh out the pattern made obvious in the last plot, using our age distribution histogram

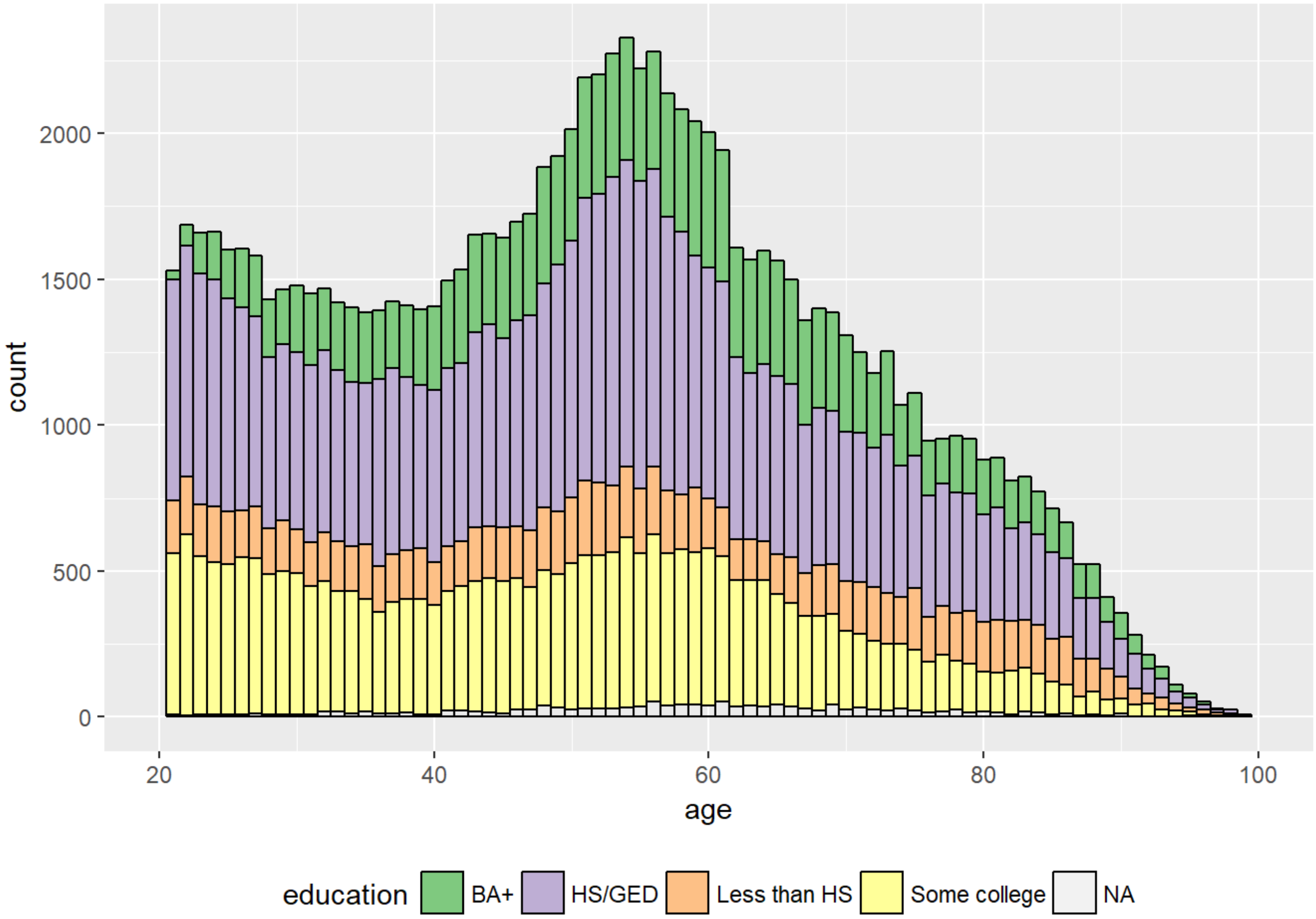




We can see how fewer females die due to guns than males, and that the peak of suicides observed in the early 20s amongst the male populace is missing amongst the females. I have not considered rates here as the ratio of males to females is 1.06 in the US, so the raw number speak for themselves. A more side by comparison of male and female statistics may help the comparison a bit more.

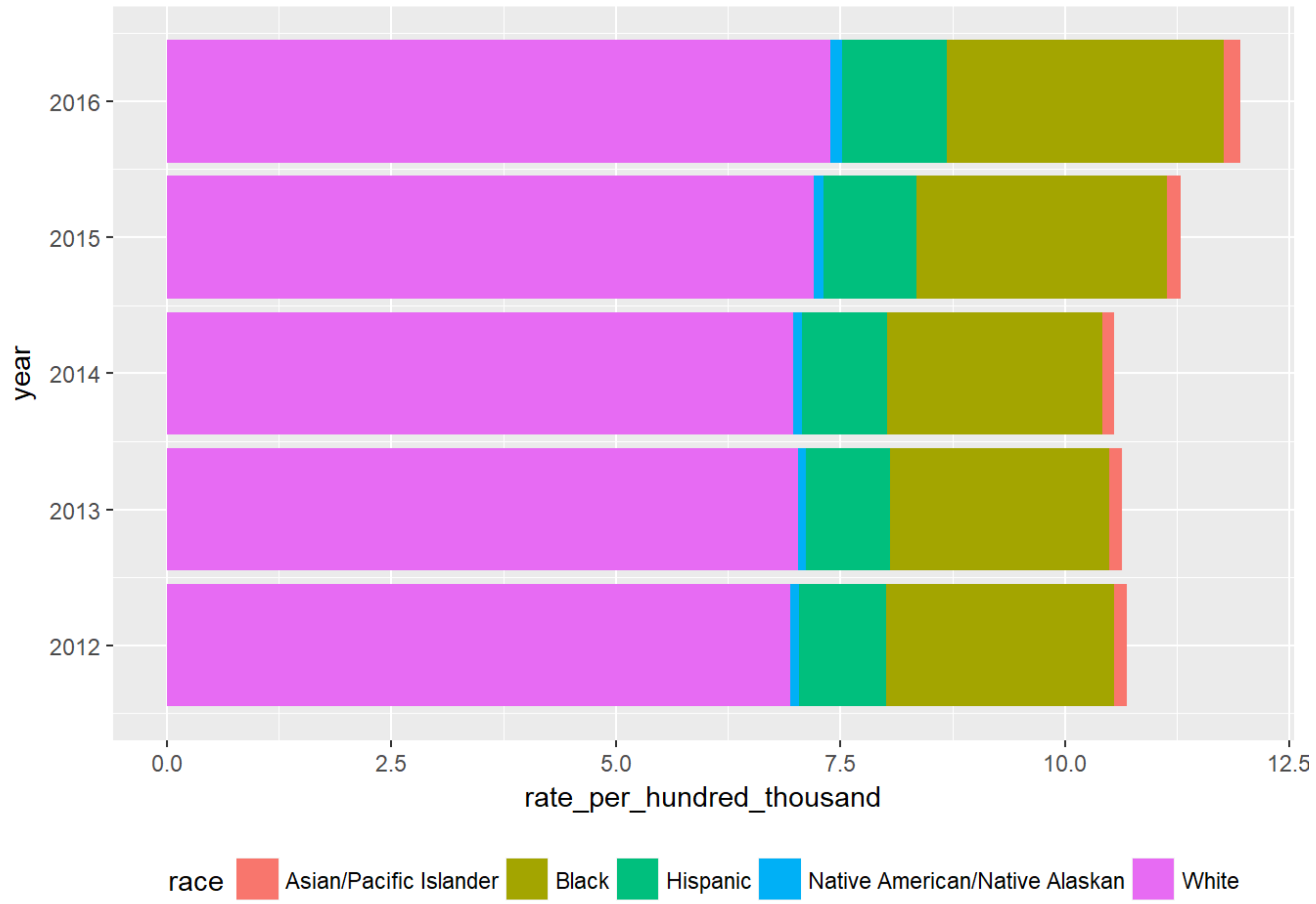


In the suicide peak, it may be worth looking at the education levels of the victims. This may give us insight into the effect that education has on suicide via guns. It is important to note that education itself is a factor of age, and the younger cohort is unlikely to be educated beyond high school. So I will only look at the age range of 20 to 100.



We notice that the majority of people have a high school education, followed by ‘Some College’ followed by BA+. This, however, may indicate absolute numbers. The percentage of people with different levels of education in the population itself would need to be factored in, in order to make a better comparison.

The next task is to divide gun deaths by race.

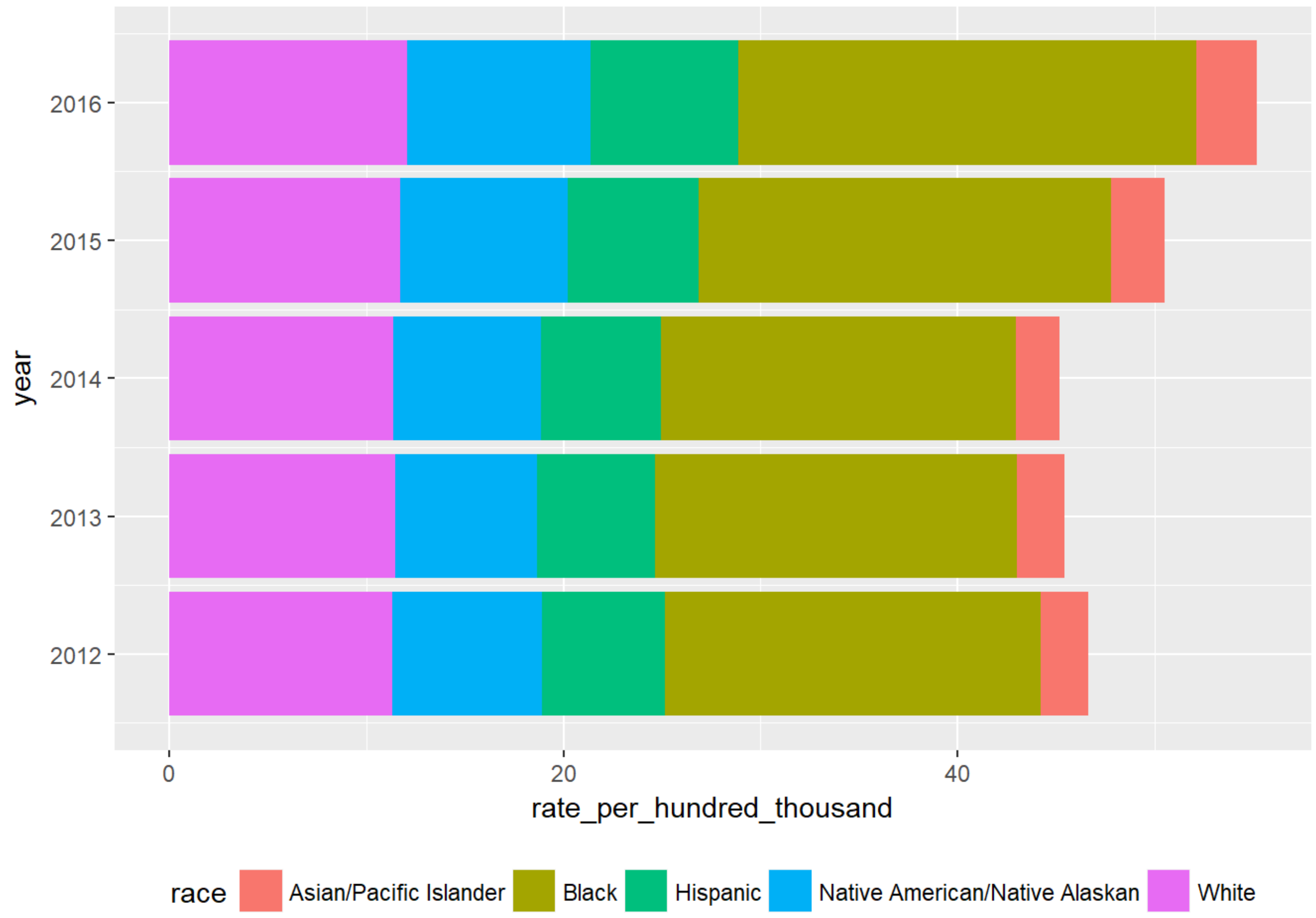


The plot above shows a few important thing. Mainly, most of the victims are white, followed by blacks, Hispanics, Asians/Pacific Islanders and Native Americans.To get a better idea of this distribution, we need to control for the representation of these groups in the population. According to the cencus bureau, as of July 1st 2016,

```
##                             race perc
## 1                          White 61.3
## 2                          Black 13.3
## 3                       Hispanic 15.6
## 4         Asian/Pacific Islander  5.9
## 5 Native American/Native Alaskan  1.3
## 6                          Other  2.6
```

I am going to use these figures universally, even though there have been marginal changes in the last few years. Therefore, we must create a new data frame to reflect these numbers

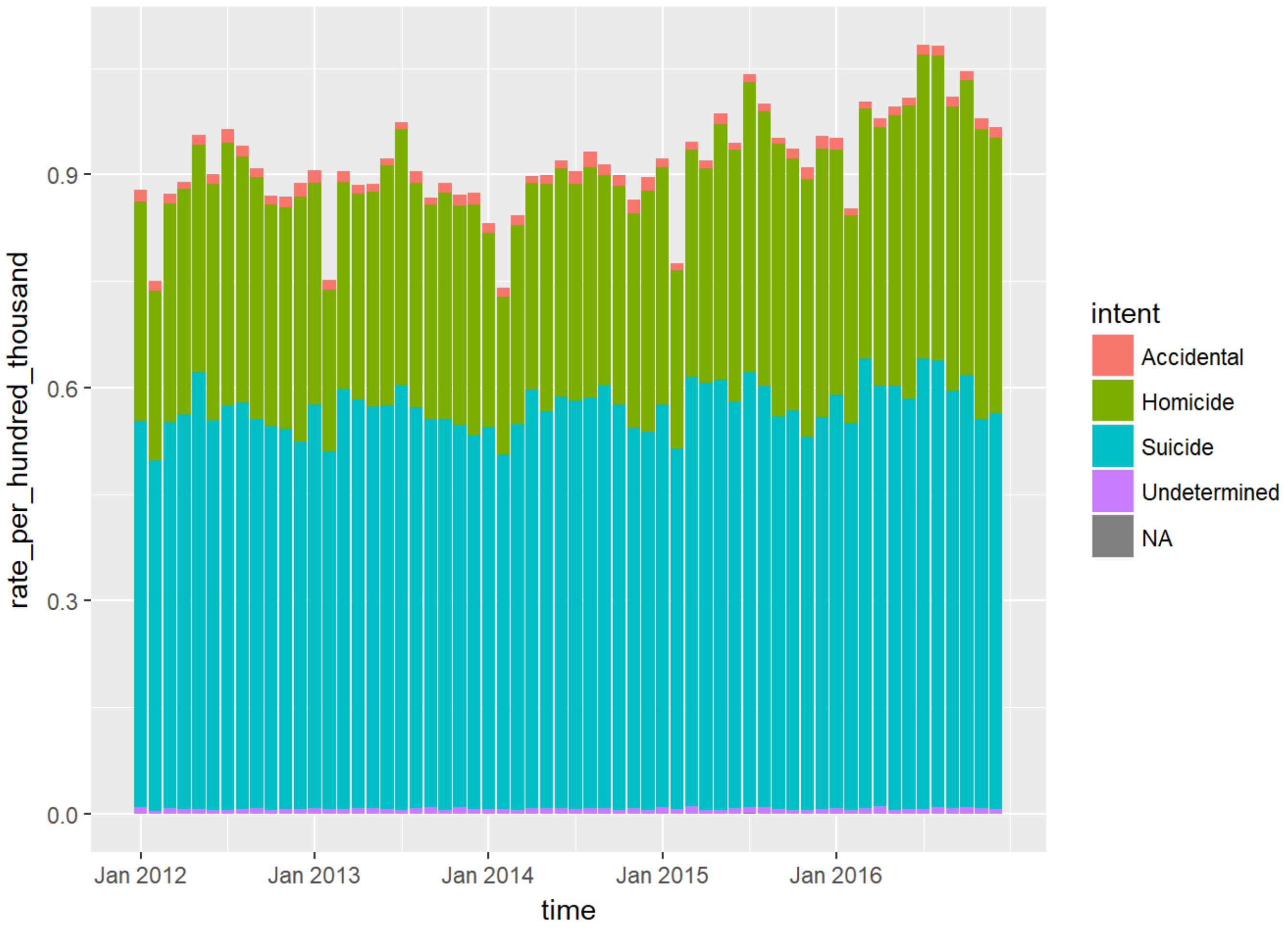
U



Adjusting for the percentage of these races in the population we see that people of African American descent are much more likely to be victims of gun violence than any other race. I will creat one final data frame that groups all important variables and gives us counts and populations for each of the relevant factors for further analysis

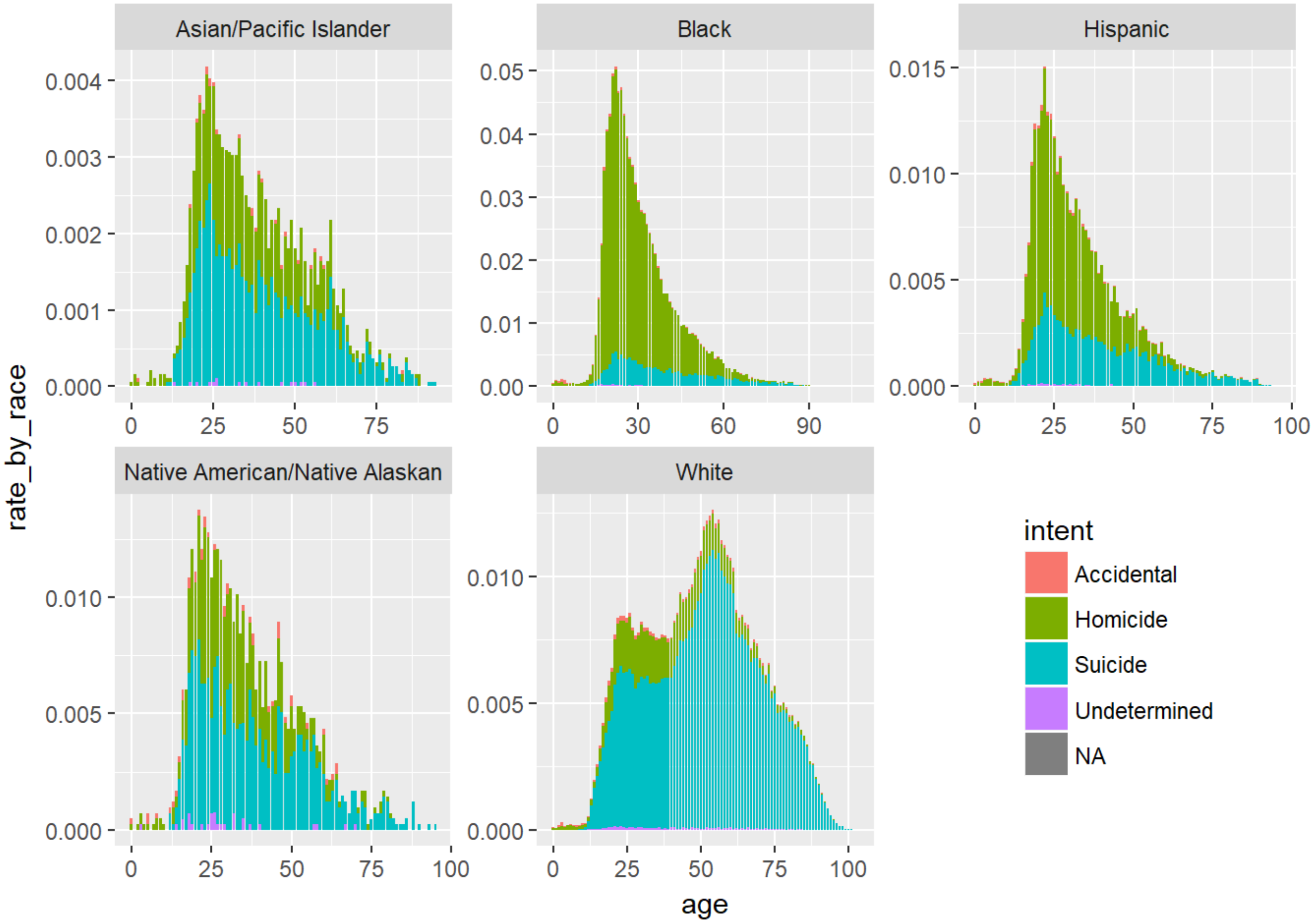
```
## # A tibble: 70,097 x 11
## # Groups: time, year, age, sex, race, intent, police [39,666]
##      time  year   age sex    race  inte~ poli~ place count rate_p~ rate_b~
##     <dbl> <dbl> <int> <fctr> <chr> <fct> <fct> <fct> <int>   <dbl>   <dbl>
##  1 0.0833  2012     0 M      Black Homi~ 0     Home      1 3.18e-4 2.39e-5
##  2 0.0833  2012     1 F      Hisp~ Homi~ 0     Home      1 3.18e-4 2.04e-5
##  3 0.0833  2012     2 F      Black Homi~ 0     Home      1 3.18e-4 2.39e-5
##  4 0.0833  2012     2 M      Hisp~ Homi~ 0     Home      1 3.18e-4 2.04e-5
##  5 0.0833  2012     3 M      Hisp~ Homi~ 0     Home      1 3.18e-4 2.04e-5
##  6 0.0833  2012     6 F      White Homi~ 0     Home      1 3.18e-4 5.20e-6
##  7 0.0833  2012     7 M      White Homi~ 0     Home      1 3.18e-4 5.20e-6
##  8 0.0833  2012     9 M      White Homi~ 0     Home      1 3.18e-4 5.20e-6
##  9 0.0833  2012    11 F      Black Suic~ 0     Home      1 3.18e-4 2.39e-5
## 10 0.0833  2012    11 M      Black Homi~ 0     Home      1 3.18e-4 2.39e-5
## # ... with 70,087 more rows
```

Now we can look at the rate of gun deaths with time colored by intent so that we get a better sense of changes in the types of gun deaths over time.



A fact that seems to jump out looking at this plot is that the increase in gun deaths that we observed earlier are mostly due to homicides. This could be further explored if we extracted all the homicide cases from the CDC data, and drew a comparison over time.

Now we can subdivide the intents by age and race



the rate of death is slightly higher for people of Hispanic and African American descent, while it is fairly low for Asians, Hawaiians, and other Pacific Islanders. Subverting our earlier finding about suicides being a bigger problem than homicides in terms of gun deaths, we see that in the hispanic and African American Communities, homicides are infact higher than suicides.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the
investigation. Were there features that strengthened each other in terms of
looking at your feature(s) of interest?

I found That the relationship between age, gender and the intent of gun death are pretty strong. Most gun victims are male, among whom the younger cohort in their 20s are victims of homicide, while the older are victims of suicide.

Were there any interesting or surprising interactions between features?

I did not find any interesting relationships between features.

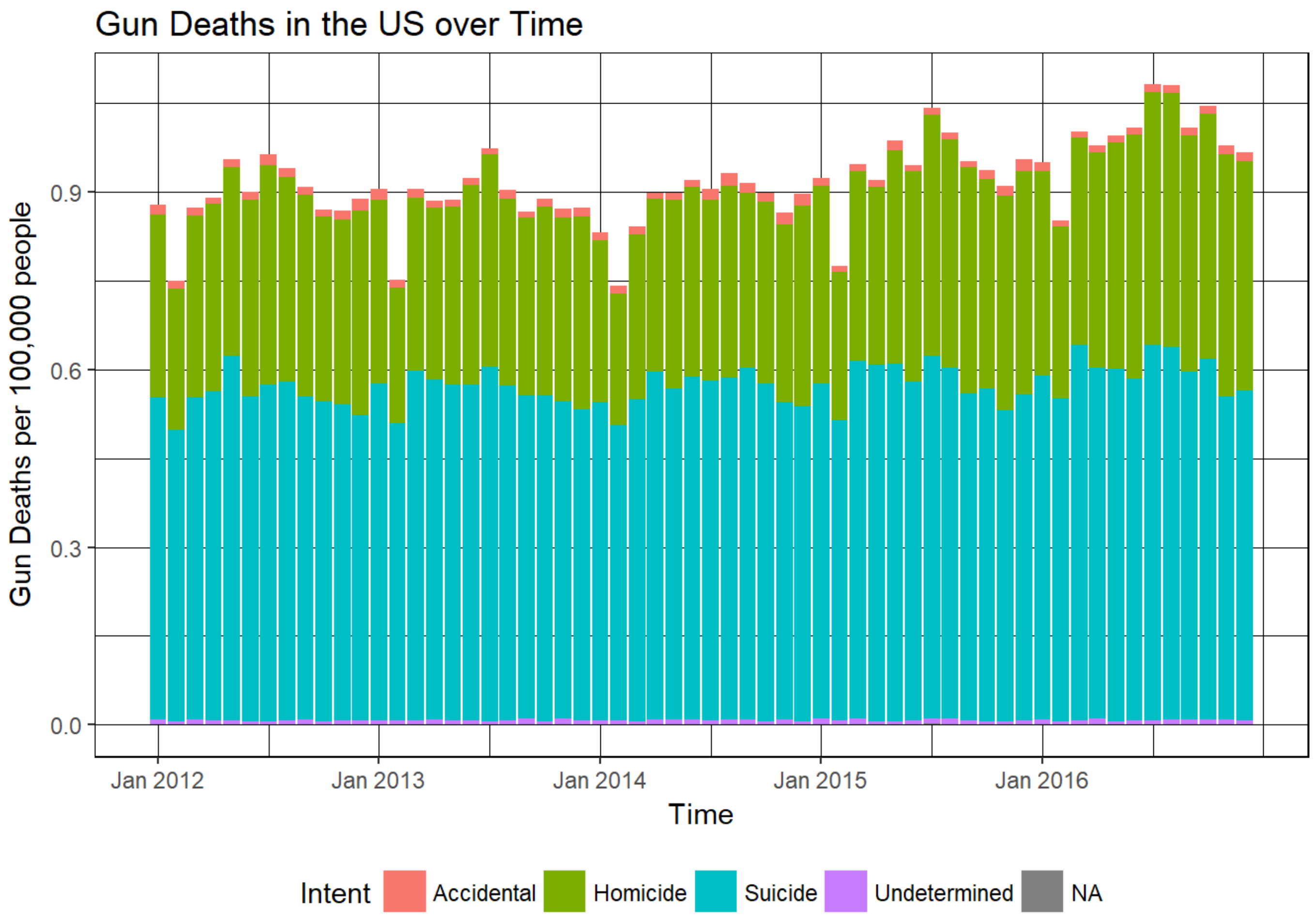
OPTIONAL: Did you create any models with your dataset? Discuss the

strengths and limitations of your model.

I did not create a model with my dataset. I am unaware of what such a model would ty to predict, or what it would look like. I suppose it should try to predict the likelihood of a person dying of gun violence, given their race, sex, age, and education level. It could also tries to predict the anticipated number of gun deaths in the US for a given year, given the demographic information about the population's race, sex, age and level of education.

Final Plots and Summary

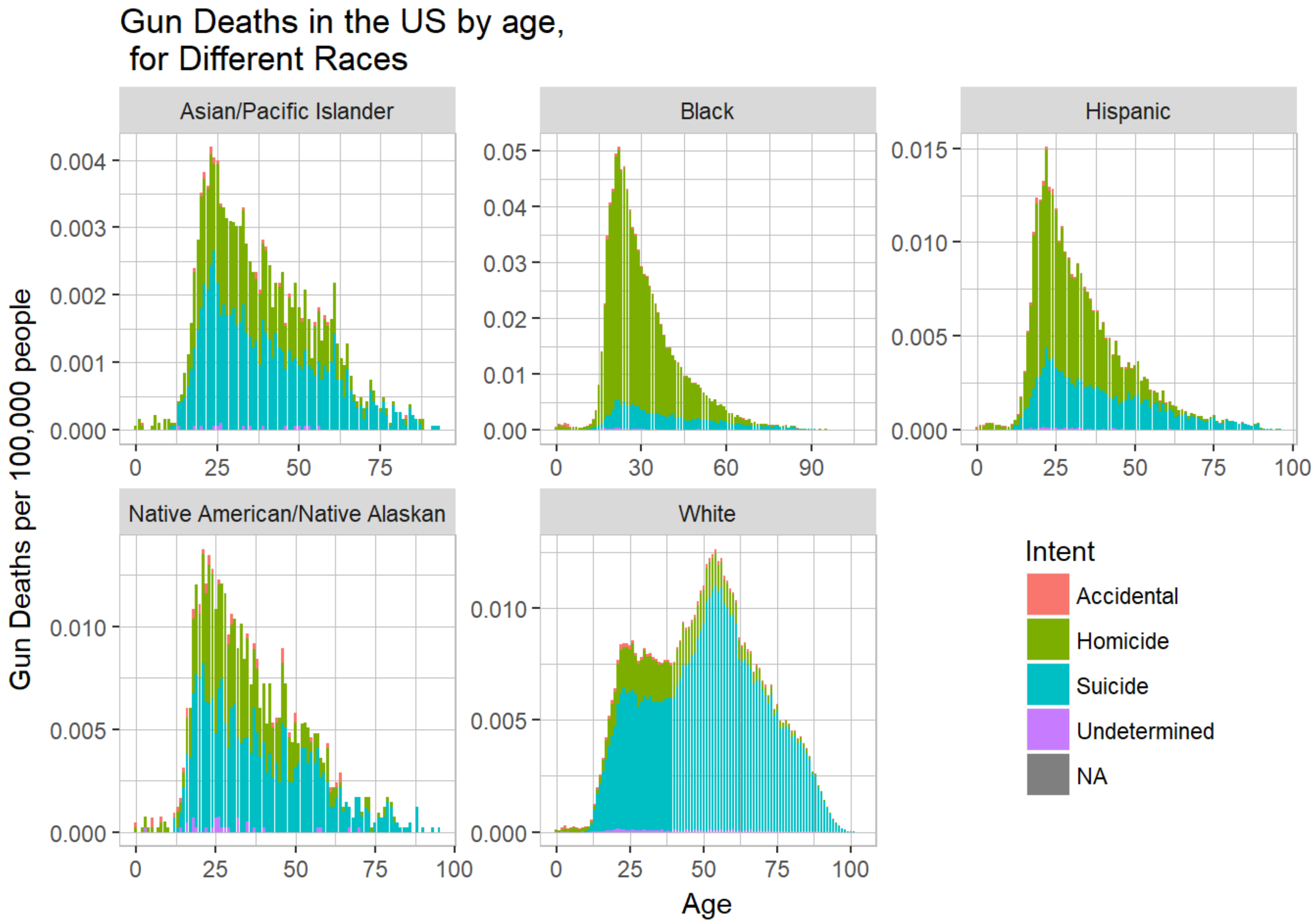
Plot One



Description One

The plot above shows that the rate of gun deaths in the US in the last 5 years while varying month to month, is gradually increasing over time. Every year seems to see a dip in the number of gun deaths in February, ad a peak in July. The former is probably due to February having fewer days. I have not been able to find a convincing explanation of the latter.

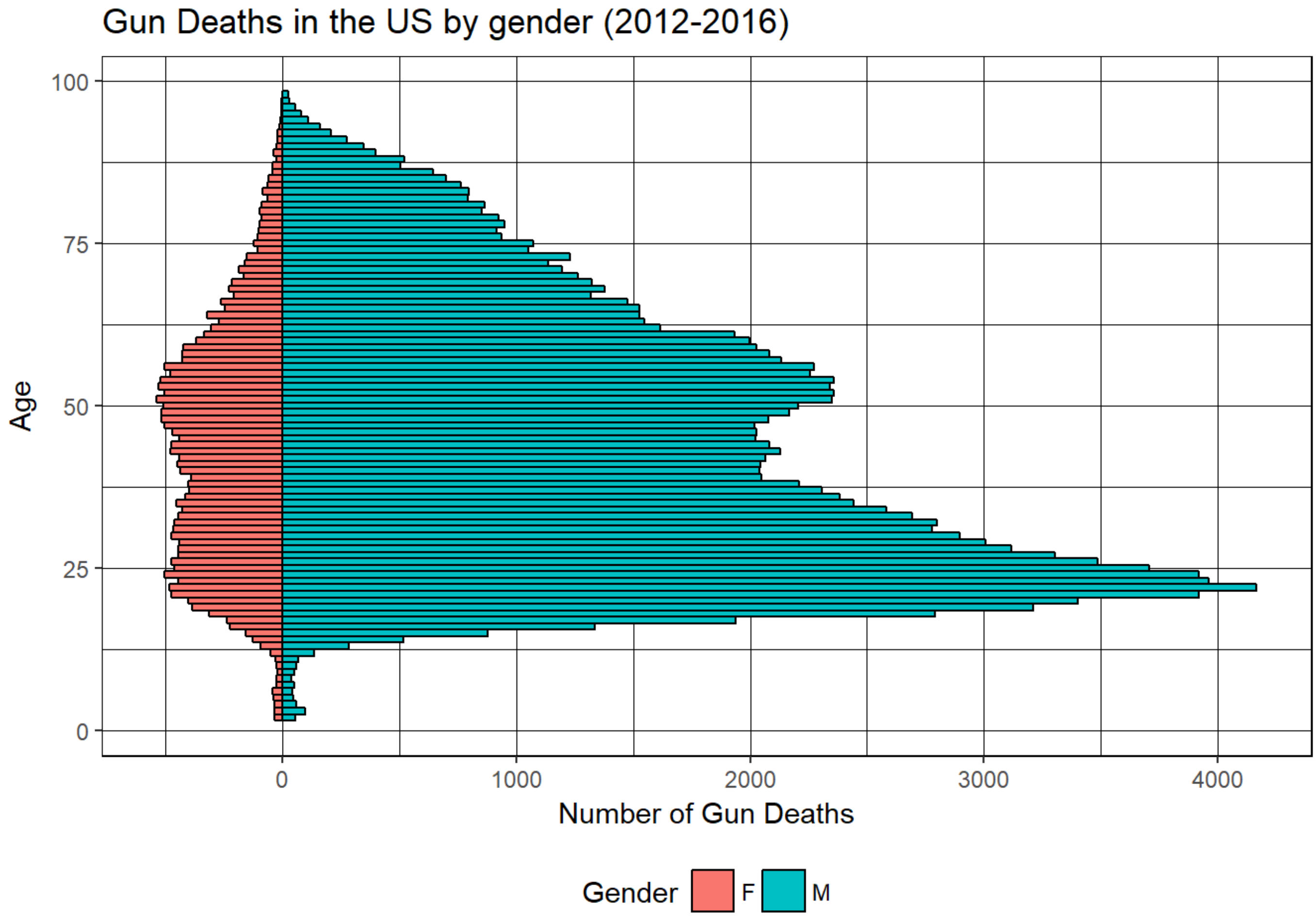
Plot Two



Description Two

The plot above shows the rate of gun death per 100,000 people for every race, colored by the intent. The implications of these plots are dire and untasteful. To spell out some of them, we can clearly see the marks of the proverbial quarter-life and mid-life crises, in the spike in gun suicides in the early 20s and 50s, although they can only be really fleshed out if we looked at the data for all suicides in the country. Moreover, the rate of death is slightly higher for people of Hispanic and African American descent, while it is fairly low for Asians, Hawaiians, and other Pacific Islanders.

Plot Three



Description Three

I shall leave the reader with this final plot of male vs female gun deaths by age. This sheds light on how the rate of gun death for men is nearly eight times what it is for women. The peaks in the mid-twenties and fifties show up for both genders but are a lot more pronounced for men than for women.

Reflection

This has been a grim and emotionally taxing exploration, albeit an important one. The primary insights are as follows:

- While the media and the American diaspora devolve into an ugly debate about gun control every time there is a mass shooting, the data shows us that the bigger issue when it comes to gun deaths is that of suicides, happening almost evenly, across time. While controlling access to guns will not take care of suicidal tendencies in the population, there is a finality to firearm suicides that doesn't exist with other methods. This point can be better made by exploring rates of suicide across several countries
- We found that the quarter-life and mid-life crises so excessively talked about in popular culture are ferociously real, as evidenced by the patterns found in the suicide data
- We observed that the rate of gun deaths, across both homicides and suicides, are much lower for women, compared to men.
- finally, we learned that the rate of gun deaths in the US seems to be rising, and if we are ever going to do anything about gun control, the time is now.

Some of the struggles in working with the data were the severely twisted syntax that ggplot often makes one go through for things that one would assume simply. For example, the pyramid chart shown in the last section should have included a label on each side saying “Female” and “Male” Respectively, so that the color parameter could be used for other variables. This, however, proved excruciatingly difficult, and after several tries, I gave up. The massively categorical nature of the data did make it easy to see patterns and that did make the exploration go well.

Further exploration of this dataset could yield useful models for public health purposes. For example -

- predict the likelihood of a person dying of gun violence, given their race, sex, age, and education level.
- predict the anticipated number of gun deaths in the US for a given year, given the demographic information about the population’s race, sex, age and level of education.

As the macabre denouement of this exploration I would like to leave the reader with this musing - would the ominous manifestation of the universality of these existential crises shown here give pause to the people suffering from these pathologies? And if they were to find solace in the realization that in fact all of us go through these points in life, would it disrupt this very pattern?