Analysis of the Washington Post Fatal Force Database

Jack Liddicoat

2024-06-24

The packages we will need to load in for this analysis:

```
library(ggplot2)
library(tidyverse)
library(marginaleffects)
library(patchwork)
library(lubridate)
library(mgcv)
library(readr)
library(usmap)
library(stargazer)
library(multcomp)
theme_set(theme())
```

Load in the data

```
df <- read_csv("polshoot.csv")
```

```
df %>% head()
```

```
## # A tibble: 6 x 12
##   date   name               age gender armed race  city  state flee  body_camera
##   <chr>  <chr>            <dbl> <chr>  <chr> <chr> <chr> <chr> <chr> <lgl>
## 1 1/2/15 Lewis Lee Lembke    47 male   gun   White Aloha OR    not   FALSE
## 2 1/2/15 Tim Elliot          53 male   gun   Asian Shel~ WA    not   FALSE
## 3 1/3/15 John Paul Quint~    23 male   unar~ Hisp~ Wich~ KS    not   FALSE
## 4 1/4/15 Kenneth Joe Bro~    18 male   gun   White Guth~ OK    not   FALSE
## 5 1/4/15 Michael Rodrigu~    39 male   other Hisp~ Evans CO    not   FALSE
## 6 1/4/15 Matthew Hoffman     32 male   repl~ White San ~ CA    not   FALSE
## # i 2 more variables: signs_of_mental_illness <lgl>,
## #   police_departments_involved <chr>
```

```
ps_cleaned <- df %>%
  mutate(body_camera = ifelse(body_camera == "TRUE", "yes", "no")) %>%
  mutate(signs_of_mental_illness = ifelse(signs_of_mental_illness == "TRUE", "yes", "no")) %>%
  mutate(date = as.Date(date, "%m/%d/%y")) %>%
  mutate(week_no = isoweek(date))
glimpse(ps_cleaned)
```

```
## Rows: 994
```

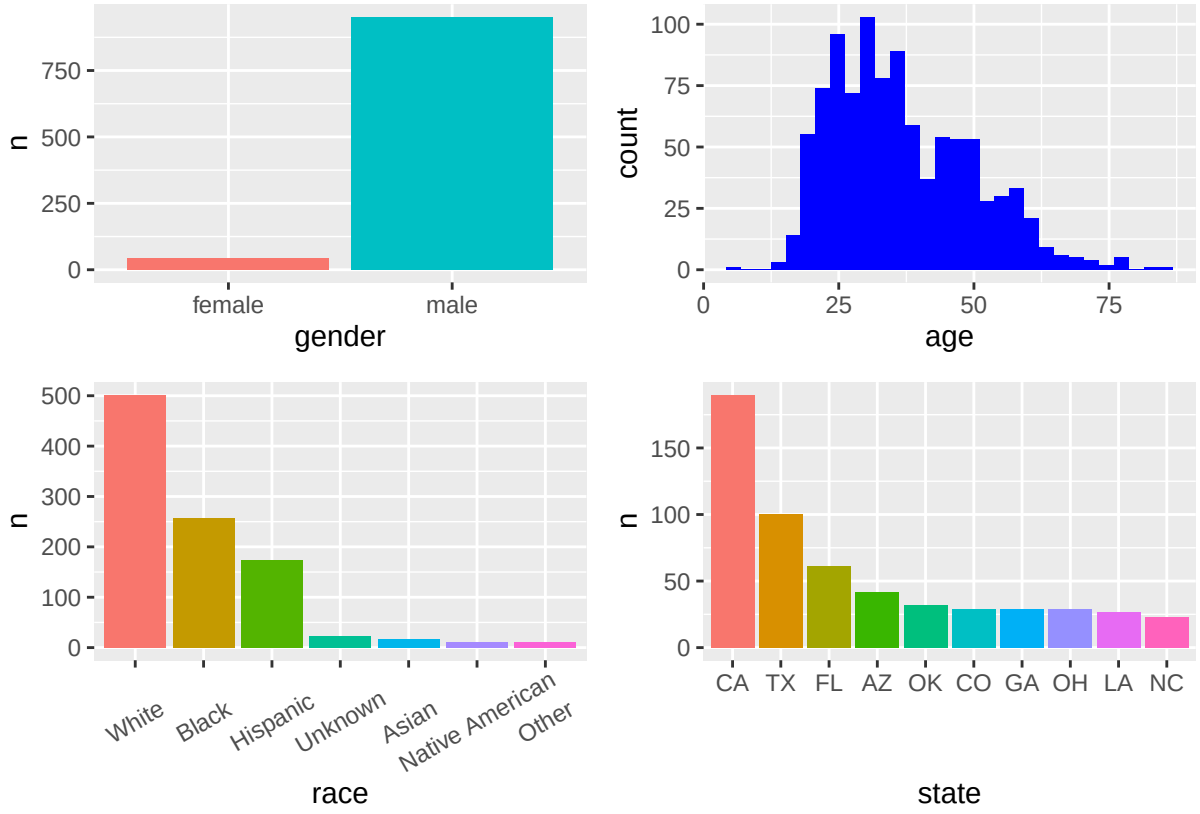
```
## Columns: 13
## $ date                          <date> 2015-01-02, 2015-01-02, 2015-01-03, 2015-~
## $ name                          <chr> "Lewis Lee Lembke", "Tim Elliot", "John Pa~
## $ age                           <dbl> 47, 53, 23, 18, 39, 32, 22, 25, 47, 34, 35~
## $ gender                        <chr> "male", "male", "male", "male", "male", "m~
## $ armed                         <chr> "gun", "gun", "unarmed", "gun", "other", "~
## $ race                          <chr> "White", "Asian", "Hispanic", "White", "Hi~
## $ city                          <chr> "Aloha", "Shelton", "Wichita", "Guthrie", ~
## $ state                         <chr> "OR", "WA", "KS", "OK", "CO", "CA", "AZ", ~
## $ flee                          <chr> "not", "not", "not", "not", "not", "not", ~
## $ body_camera                   <chr> "no", "no", "no", "no", "no", "no", "no", ~
## $ signs_of_mental_illness       <chr> "no", "yes", "no", "no", "no", "yes", "no"~
## $ police_departments_involved <chr> "Washington County Sheriff's Office, OR", ~
## $ week_no                       <dbl> 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, ~
```

Exploration

We can look at the demographic statistics and visualize them via ggplot2.

```
gender <- ps_cleaned %>%
  count(gender) %>%
  ggplot(aes(gender, n, fill = gender)) +
  geom_bar(stat = "identity", show.legend = F)
race <- ps_cleaned %>%
  count(race) %>%
  mutate(race = fct_reorder(race, -n)) %>%
  ggplot(aes(race, n, fill = race)) +
  geom_bar(stat = "identity", show.legend = F) +
  theme(axis.text.x = element_text(hjust = .5, angle = 30))
age <- ps_cleaned %>%
  ggplot() +
  geom_histogram(aes(age), fill = "blue")
states <- ps_cleaned %>% # just going to pick the 10 most highest
  count(state) %>%
  arrange(desc(n)) %>%
  head(10) %>%
  mutate(state = fct_reorder(state, -n)) %>%
  ggplot(aes(state, n, fill = state)) +
  geom_bar(stat = "identity", show.legend = F)

gender + age + race + states
```

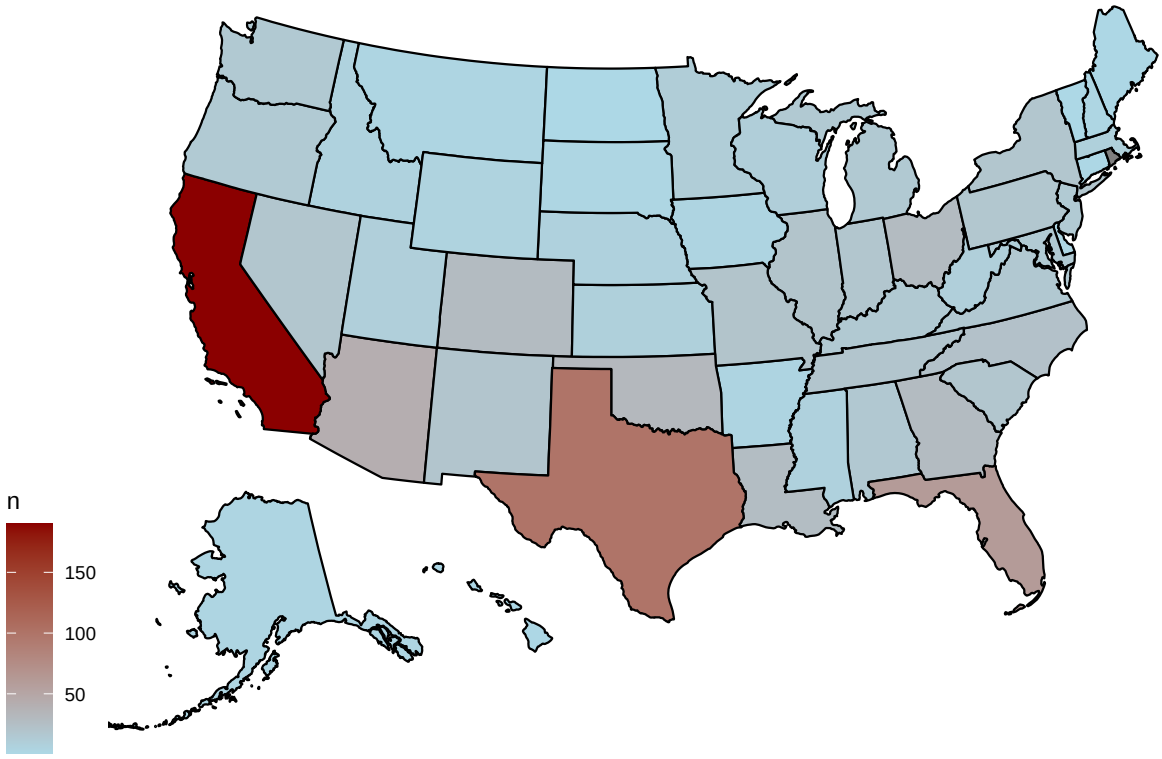
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can conclude that most people who were shot and killed by the police in 2015 were male, they skewed younger in age, they were mostly white (though adjusted for population size were more likely to be black or Hispanic), and most shootings occured in the most populus states, with some exceptions (e.g., Arizona, Oklahoma, Colorado).

We can make a map of the U.S. to better visualize the distribution of police killings. Note that this is **not per-capita**, so it largely reflects differences in population.

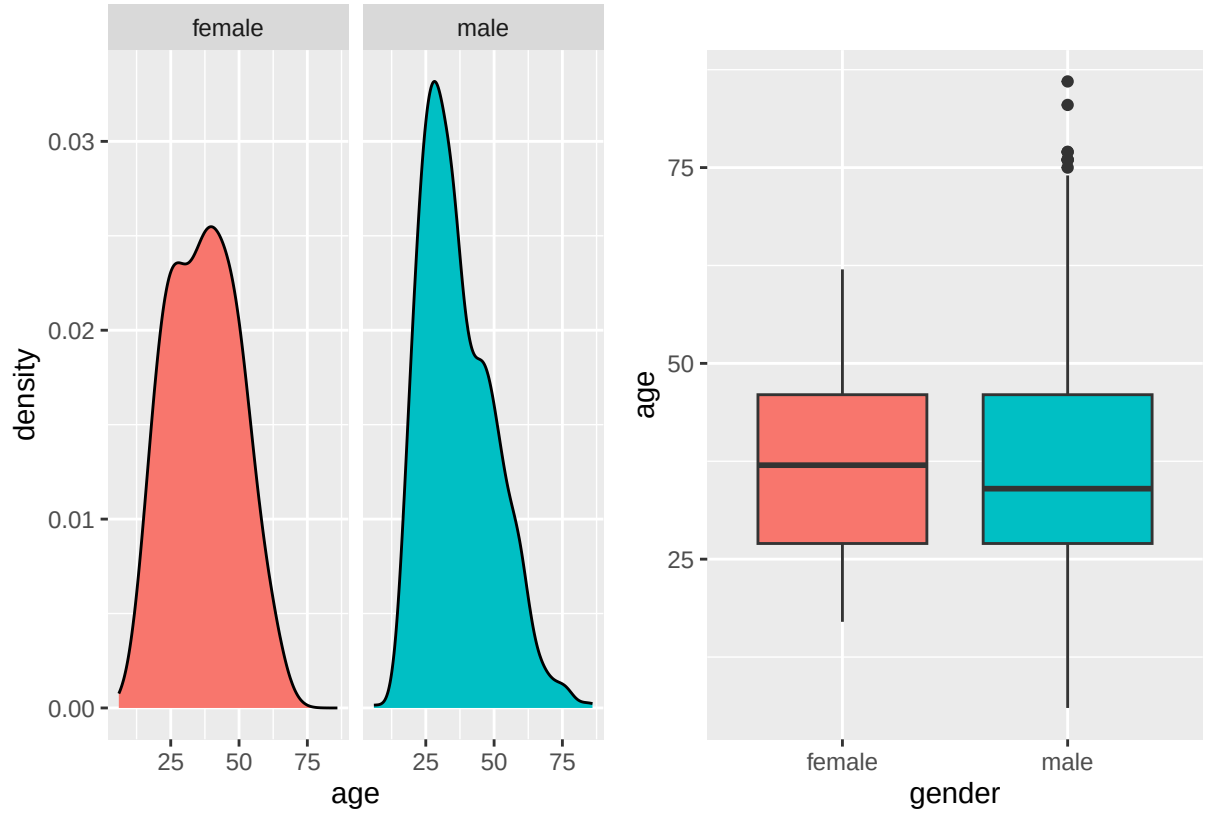
```
map_data <- ps_cleaned %>%
  count(state)
plot_usmap(regions = "state", values = "n", data = map_data) +
  scale_fill_continuous(low = "lightblue", high = "darkred")
```



From here, we can ask more questions. How does race interact with age? How does gender interact with age? Are there differences in the circumstances of police shootings for these different variables?

Here is what the plots of gender and age looks like.

```
gender_density <- ps_cleaned %>%
  ggplot() +
  geom_density(aes(age, fill = gender), show.legend = F) +
  facet_wrap(~gender)
gender_box <- ps_cleaned %>%
  ggplot() +
  geom_boxplot(aes(gender, age, fill = gender), show.legend = F)
gender_density + gender_box
```



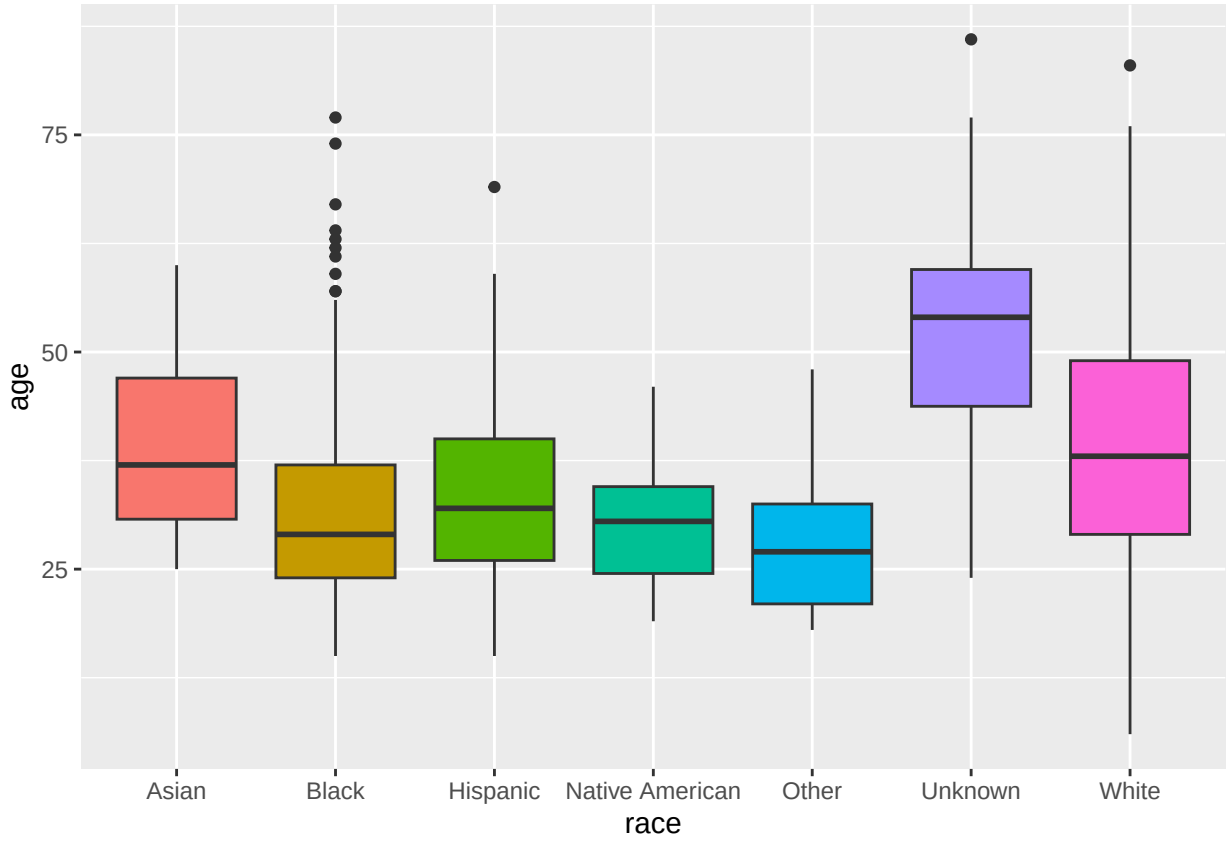
It does appear as if men tend to skew a bit younger than women. We can check this using some summary statistics.

```
tbl <- ps_cleaned %>%
  filter(!is.na(gender), !is.na(age)) %>%
  group_by(gender) %>%
  summarise(`25th Percentile` = quantile(age, .25),
            Median = median(age),
            `75th Percentile` = quantile(age, .75))
tbl
```

```
## # A tibble: 2 x 4
##   gender `25th Percentile` Median `75th Percentile`
##   <chr>              <dbl>  <dbl>             <dbl>
## 1 female                27     37                46
## 2 male                  27     34                46
```

They are very similar in age. We can now see if different racial groups also have similar age distributions.

```
ps_cleaned %>%
  ggplot() +
  geom_boxplot(aes(race, age, fill = race), show.legend = F)
```



On its face, it seems like whites and Asians are significantly older than blacks, Hispanics, and Native Americans. We can do an ANOVA to test if there is a significant difference in age between the racial groups.

```
ps_cleaned$race <- as.factor(ps_cleaned$race)
res_aov <- aov(age ~ race, data = ps_cleaned)
summary(res_aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## race          6  18791  3131.8   20.69 <2e-16 ***
## Residuals   979 148211   151.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 8 observations deleted due to missingness
```

From the test, we can see that our F-statistic is highly significant ($F = 20.69$, $p < .001$). Hence, we reject the null hypothesis that the age of the victims is the same between the 6 groups, meaning at least one group is different. However, we should check if that still holds true assuming unequal variances. We will proceed with Welch's test:

```
oneway.test(age ~ race,
  data = ps_cleaned,
  var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
```

```
##
## data:  age and race
## F = 17.981, num df = 6.000, denom df = 47.646, p-value = 9.196e-11
```

Our F-statistic dropped a tad bit (from 20.69 to 17.98), but we still reject the null hypothesis stated above. We can see the difference between pairs if we run a Tukey HSD test.

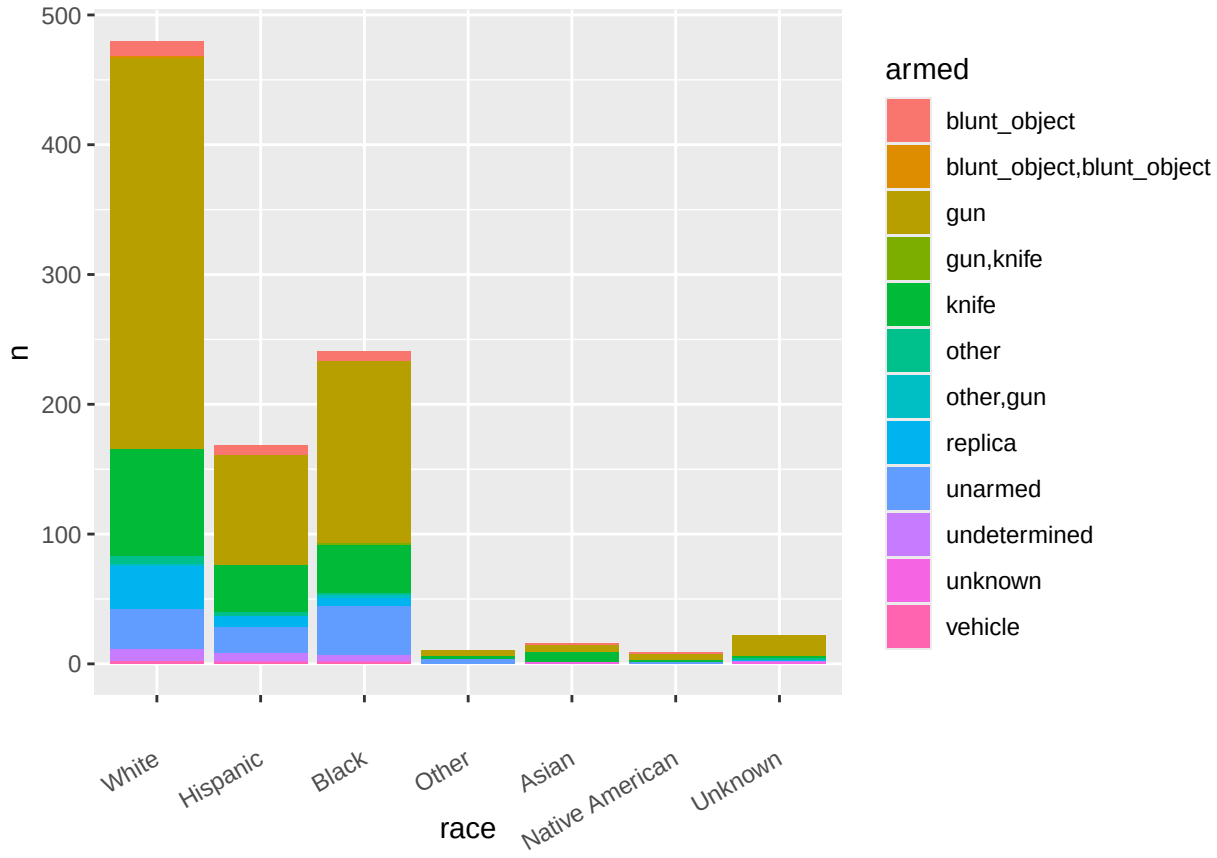
```
tukey_test <- glht(res_aov,
  linfct = mcp(race = "Tukey")
)
summary(tukey_test)
```

```
## Warning in RET$pfunction("adjusted", ...): Completion with error > abseps

## Warning in RET$pfunction("adjusted", ...): Completion with error > abseps

## Warning in RET$pfunction("adjusted", ...): Completion with error > abseps

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = age ~ race, data = ps_cleaned)
##
## Linear Hypotheses:
##                                   Estimate Std. Error t value Pr(>|t|)
## Black - Asian == 0                -7.63284    3.17105  -2.407   0.1600
## Hispanic - Asian == 0             -6.05853    3.21511  -1.884   0.4333
## Native American - Asian == 0      -9.22500    4.95992  -1.860   0.4489
## Other - Asian == 0               -11.02500    4.95992  -2.223   0.2376
## Unknown - Asian == 0              14.17500    4.12691   3.435    <0.01 **
## White - Asian == 0                 0.03635    3.12465   0.012   1.0000
## Hispanic - Black == 0              1.57432    1.21193   1.299   0.8184
## Native American - Black == 0      -1.59216    3.96644  -0.401   0.9995
## Other - Black == 0                -3.39216    3.96644  -0.855   0.9719
## Unknown - Black == 0              21.80784    2.85713   7.633    <0.01 ***
## White - Black == 0                 7.66920    0.94618   8.105    <0.01 ***
## Native American - Hispanic == 0   -3.16647    4.00176  -0.791   0.9810
## Other - Hispanic == 0             -4.96647    4.00176  -1.241   0.8482
## Unknown - Hispanic == 0           20.23353    2.90596   6.963    <0.01 ***
## White - Hispanic == 0              6.09488    1.08474   5.619    <0.01 ***
## Other - Native American == 0      -1.80000    5.50254  -0.327   0.9999
## Unknown - Native American == 0    23.40000    4.76534   4.910    <0.01 ***
## White - Native American == 0       9.26135    3.92945   2.357   0.1793
## Unknown - Other == 0              25.20000    4.76534   5.288    <0.01 ***
## White - Other == 0                11.06135    3.92945   2.815   0.0574 .
## White - Unknown == 0             -14.13865    2.80554  -5.040    <0.01 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

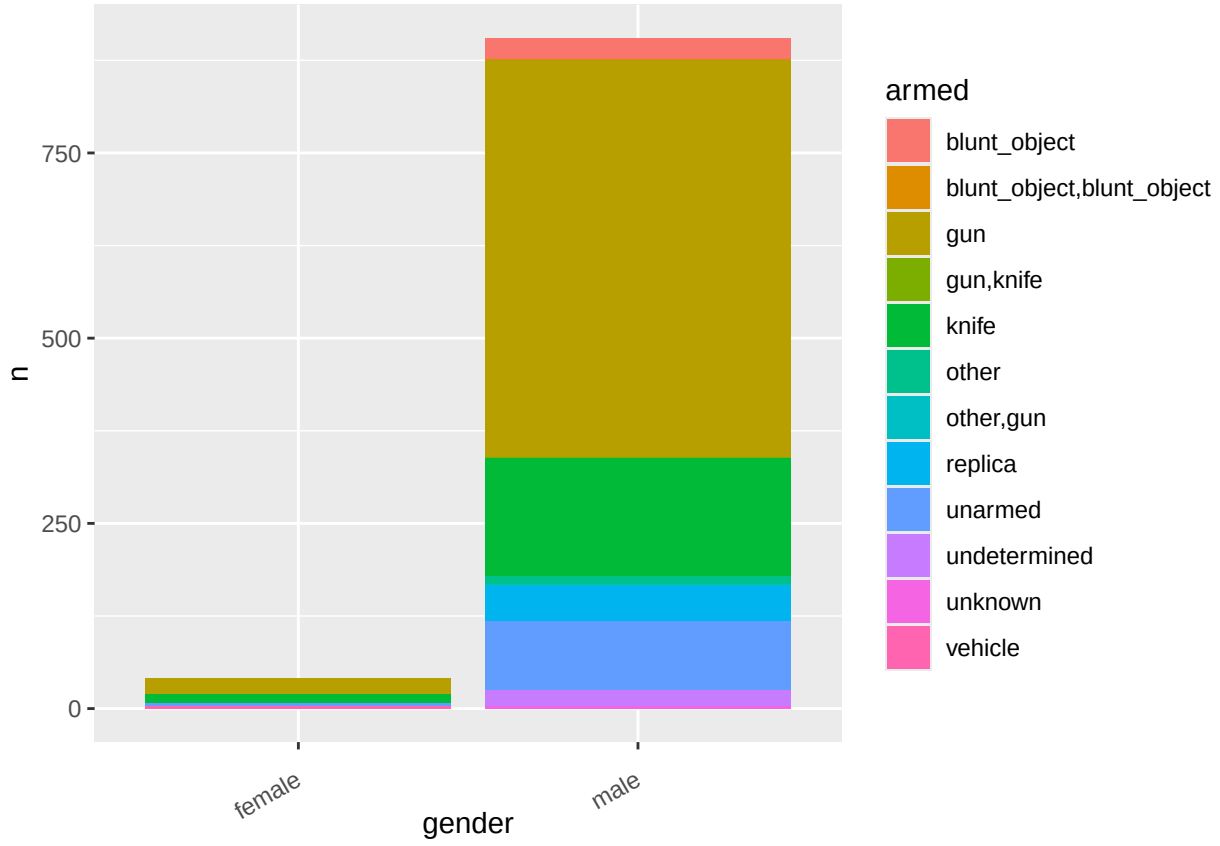
There is a lot of stuff to go through here. However, the linear hypothesis tells us, rather intuitively, that we are assuming the difference between the age of groups are the same, our t-scores and corresponding p-values inform us to fail to reject or reject that hypothesis. A few things of note: white victims are nearly 8 years older than black victims and roughly 6 years older than Hispanic victims, there is no statistically significant difference between white and Asian victims.

Lets look at the circumstances of the shootings by race. Firstly, lets look at who the distrubutions of weapons by race, age, and sex.

```
ps_cleaned %>%
  filter(!is.na(armed)) %>%
  count(race, armed, sort = T) %>%
  mutate(race = fct_reorder(race, -n)) %>%
  ggplot(aes(race, n, fill = armed)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```

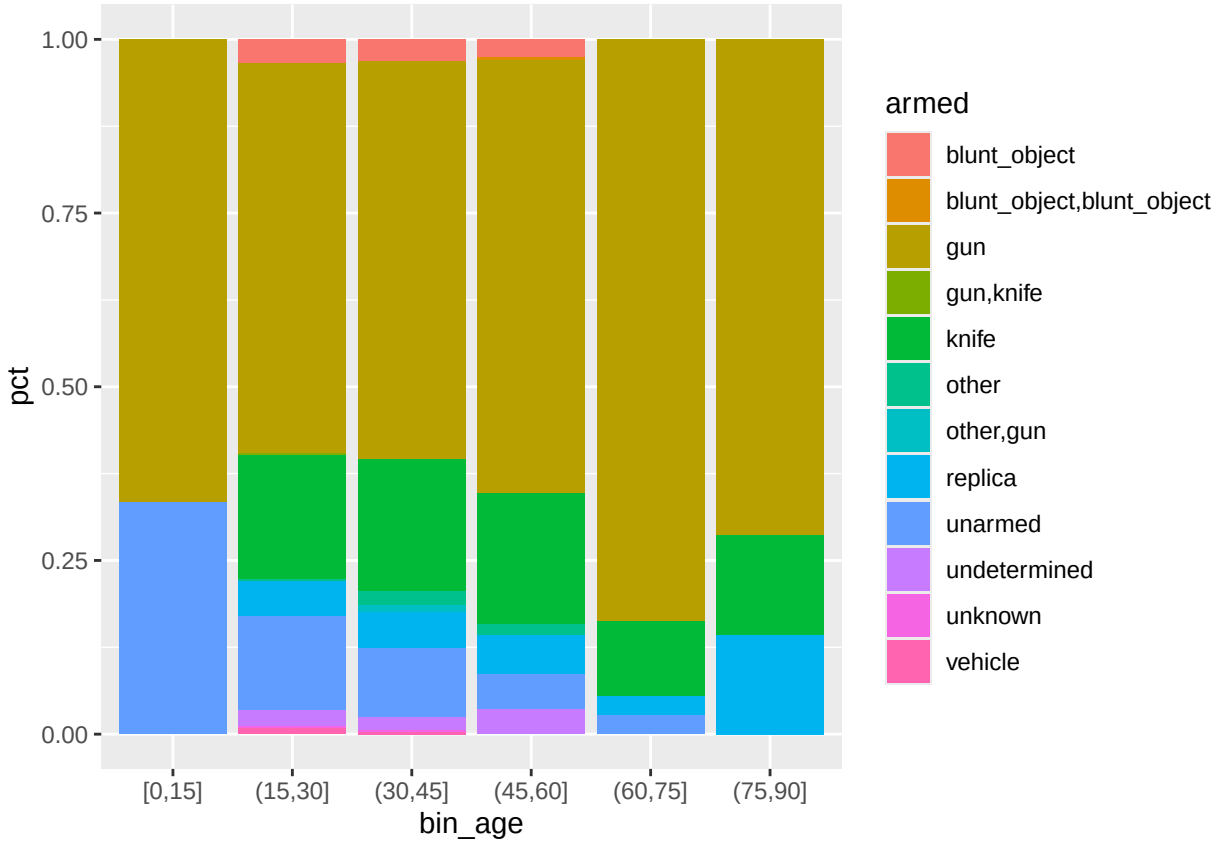



```
ps_cleaned %>%
  filter(!is.na(armed)) %>%
  count(gender, armed) %>%
  ggplot(aes(gender, n, fill = armed)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



It may be more helpful to make an area plot for the age variable.

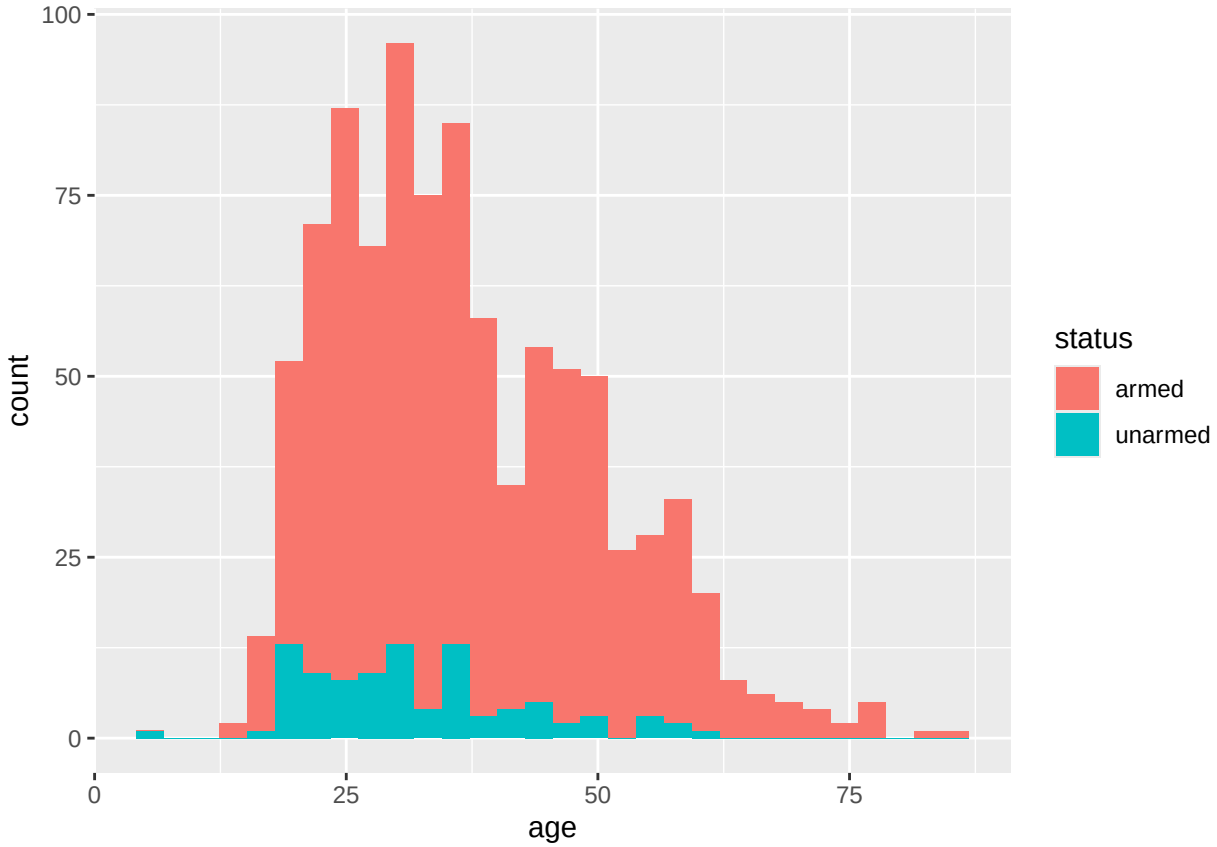
```
ps_cleaned %>%
  filter(!is.na(armed), !is.na(age)) %>%
  mutate(bin_age = cut_width(age, 15, boundary = 0)) %>%
  count(bin_age, armed) %>%
  group_by(bin_age) %>%
  mutate(pct = n/sum(n)) %>%
  ggplot(aes(x = bin_age, y = pct, fill = armed)) +
  geom_bar(stat = "identity")
```



We can see that the proportion of people who were unarmed when shot and killed by the police decreases as a function of age. We can just make the armed very binary to see this more clearly.

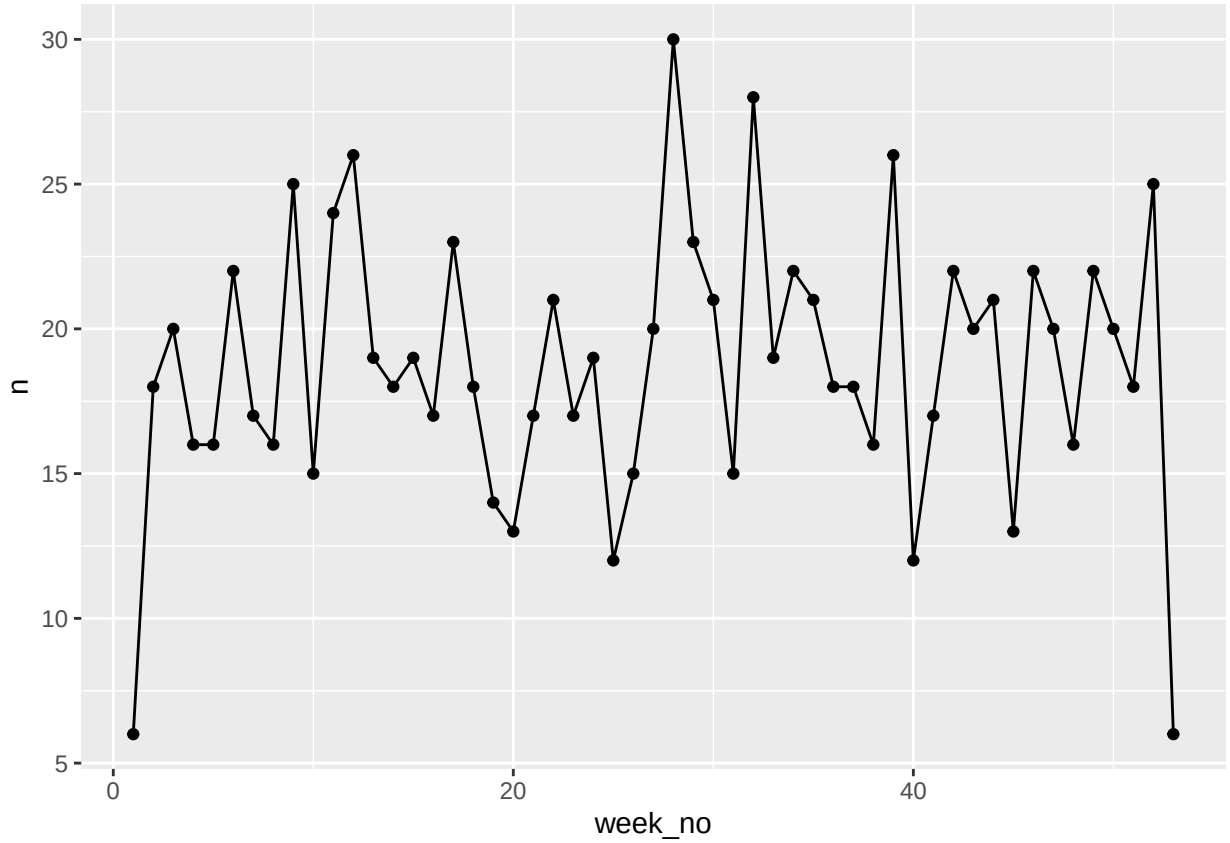
```
ps_cleaned %>%
  filter(!is.na(armed), !is.na(age)) %>%
  mutate(status = ifelse(armed == "unarmed", "unarmed", "armed")) %>%
  ggplot() +
  geom_histogram(aes(age, fill = status))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



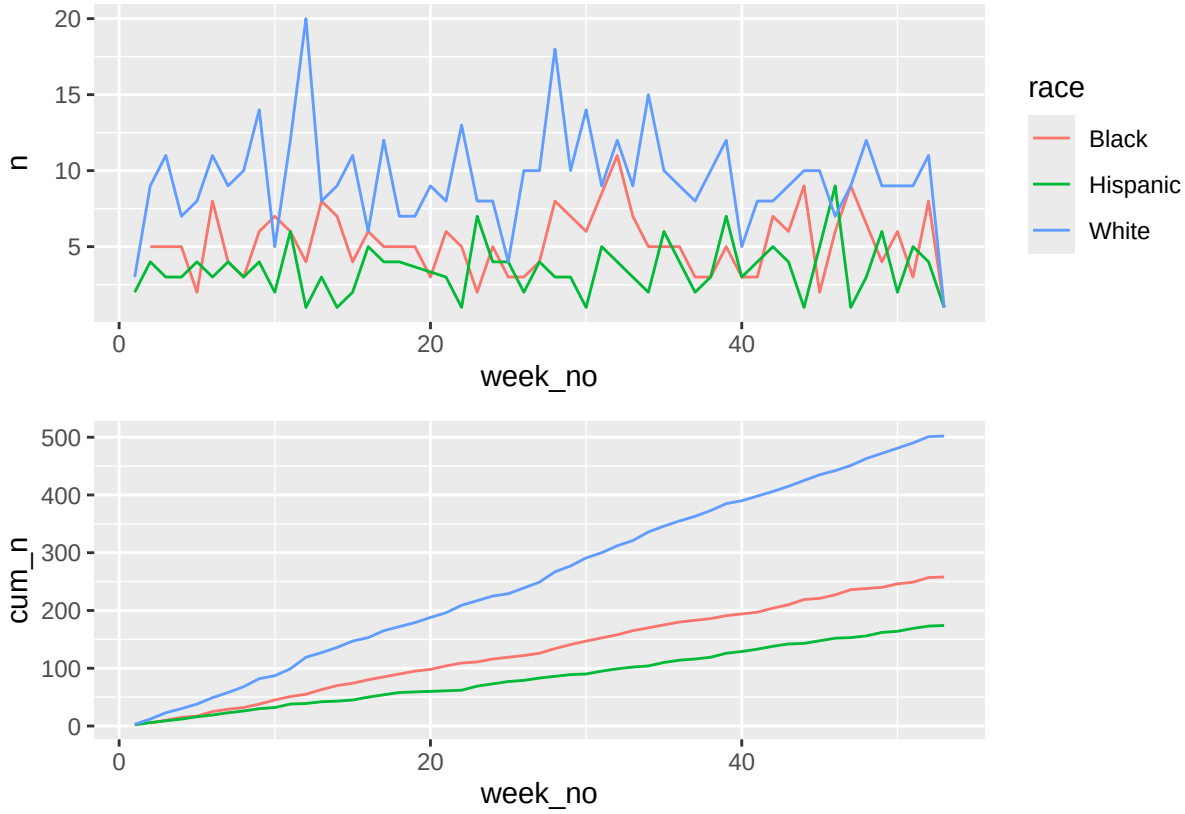
I want to know how shootings look over the course of the year. To do that, we can create a variable which groups the dates into weeks (which we have already done).

```
ps_cleaned %>%
  count(week_no) %>%
  ggplot(aes(week_no, n)) +
  geom_line() +
  geom_point()
```



As can be seen in the graph, shootings tend to be a bit higher in the summer, but the increase is not too dramatic. We can look at how this differs by race. Let's just look at whites, blacks, and Hispanics to make the graph easier to read.

```
weekly <- ps_cleaned %>%
  filter(race %in% c("Black", "White", "Hispanic")) %>%
  count(week_no, race) %>%
  group_by(race) %>%
  ggplot(aes(week_no, n, color = race)) +
  geom_line()
cumulative <- ps_cleaned %>%
  filter(race %in% c("Black", "White", "Hispanic")) %>%
  count(week_no, race) %>%
  group_by(race) %>%
  mutate(cum_n = cumsum(n)) %>%
  ggplot(aes(week_no, cum_n, color = race)) +
  geom_line(show.legend = F)

weekly / cumulative
```



Looks pretty similar for each race, although you could argue that blacks and whites experience a higher relative victimization in the summer compared to Hispanics.

Regression Analysis

For our regression analysis, we will just make the race and the gun variable binary. Since we are mostly interested in **unarmed** shootings, we can just make a binary variable for unarmed in each of these.

```
reg_data <- ps_cleaned %>%
  mutate(race = ifelse(race == "Black", 1, 0),
         unarmed = ifelse(armed == "unarmed", 1, 0),
         body_camera = ifelse(body_camera == "yes", 1, 0),
         flee = ifelse(flee == "not", 0, 1),
         signs_of_mental_illness = ifelse(signs_of_mental_illness == "yes", 1, 0))
```

For the first regression, we are going to see how race affects the probability the suspect was armed. The regression equation is as follows:

$$ln(\frac{p(unarmed)}{1-p(unarmed)}) = \beta_0 + \beta_1 race$$

```
glm1 <- glm(unarmed ~ race, data = reg_data, family = binomial)
```

```
stargazer(glm1,
          title = "Regression to Examine Effect of Race on Probability Unarmed",
          header = F)
```

Table 1: Regression to Examine Effect of Race on Probability Unarmed

	Dependent variable:
	unarmed
race	0.755^{***}
	(0.224)
Constant	-2.431^{***}
	(0.138)
Observations	946
Log Likelihood	−303.021
Akaike Inf. Crit.	610.043
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see that the coefficient is positive and statistically significant ($p < .01$). We estimate that, without controls, that blacks have 2.12 times higher odds of being unarmed when shot and killed by the police compared to non-blacks.

For our second regression equation, we can look at the effect of age on being unarmed and plot its effects using plot_predictions() from marginaleffects. Our regression equation is:

$$ln(\frac{p(unarmed)}{1-p(unarmed)}) = \beta_0 + \beta_1 age$$

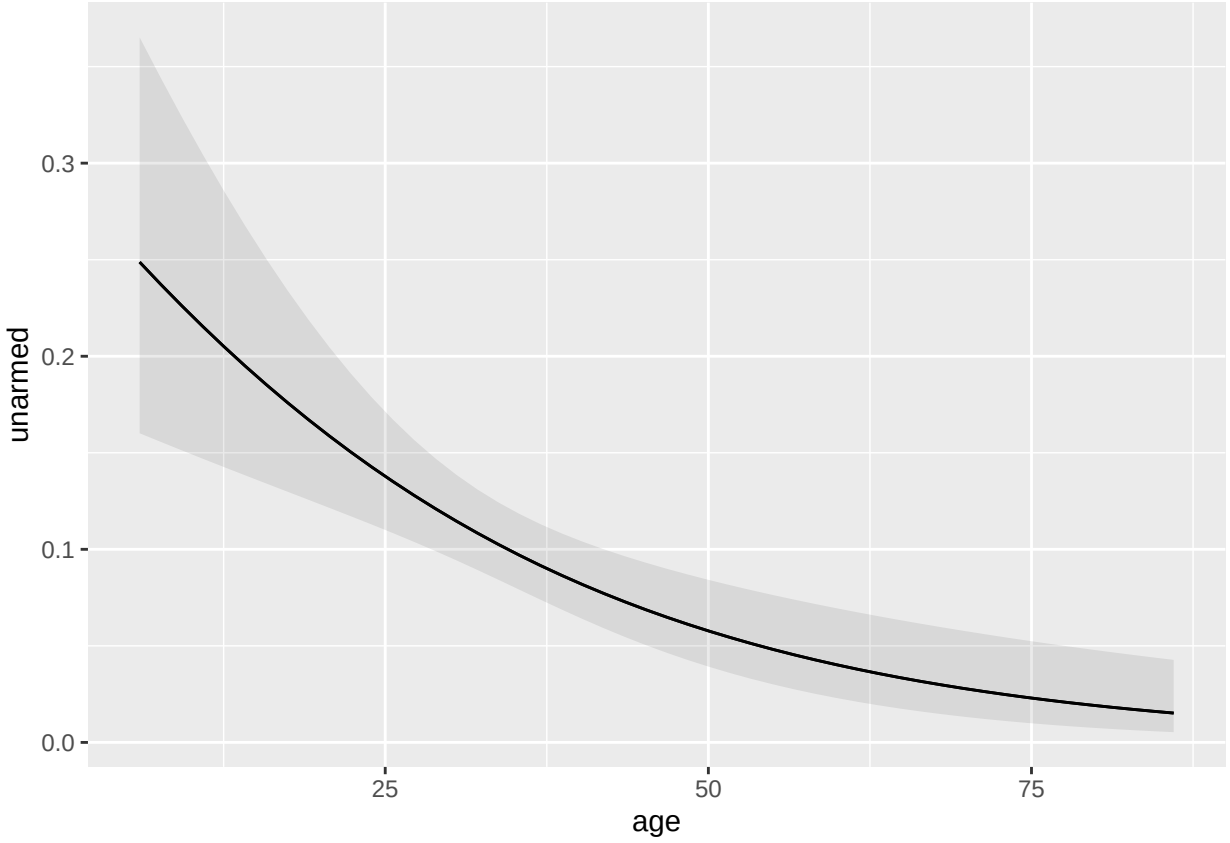
```
glm2 <- glm(unarmed ~ age, data = reg_data, family = binomial)
```

```
stargazer(glm2,
          title = "Regression to Examine Effect of Age on Probability Unarmed",
          header = F)
```

```
plot_predictions(glm2, condition = "age")
```

Table 2: Regression to Examine Effect of Age on Probability Unarmed

	Dependent variable:
	unarmed
age	-0.038^{***}
	(0.010)
Constant	-0.875^{***}
	(0.337)
Observations	938
Log Likelihood	-296.825
Akaike Inf. Crit.	597.650
Note:	*p<0.1; **p<0.05; ***p<0.01



From both the regression and the plot, we can see that age definitely decreases the probability that the victim is unarmed. The estimate from the logistic regression is that a 10-year increase in age decreases the odds that a victim is unarmed by 32%.

In the next regression, I will look at the effect of age and race on the probability that a victim was unarmed. The regression equation will be:

$$ln(\frac{p(unarmed)}{1-p(unarmed)}) = \beta_0 + \beta_1 race + \beta_2 race + \beta_2 race * age$$

To simplify the effect of age in the regression, I will center the data at the median age of a victim in the database.

```
glm3 <- reg_data %>%
  filter(!is.na(age)) %>%
  mutate(age = (age - median(age))) %>%
  glm(unarmed ~ race*age, data = ., family = binomial)
```

```
stargazer(glm3,
          title = "Regression to Examine Effect of Race and Age on Probability Unarmed",
          header = F)
```

Table 3: Regression to Examine Effect of Race and Age on Probability Unarmed

	Dependent variable:
	unarmed
race	0.679^{***}
	(0.231)
age	-0.052^{***}
	(0.013)
race:age	0.054^{***}
	(0.020)
Constant	-2.367^{***}
	(0.142)
Observations	938
Log Likelihood	−290.875
Akaike Inf. Crit.	589.751
Note:	*p<0.1; **p<0.05; ***p<0.01

The result of the regression is interesting. Controlling for age, blacks are still have 1.97 times higher odds ($p < .01$) of being unarmed when shot and killed by the police. However, we can see that, for non-blacks, a one-year increase in age decreases the odds of them being unarmed by ~5%, whereas the *opposite* is true for black victims.

The next regression uses all of the variables in the data set, besides state an city, which cause overfitting in the model, the second equation uses the interaction terms as well. The equations are:

(1)

$$ln(\frac{p(unarmed)}{1-p(unarmed)}) = \beta_0 + \beta_1 race + \beta_2 age + \beta_2 flee + \beta_3 bodycamera + \beta_4 mentallillness$$

(2)

$$ln(\frac{p(unarmed)}{1-p(unarmed)}) = \beta_0 + \beta_1 race + \beta_2 age + \beta_2 flee + \beta 3 bodycamera + \beta_4 race * age + \beta_5 race * flee + \beta_6 race * bodycamera + \beta$$

```r
glm4 <- reg_data %>%
  filter(!is.na(age)) %>%
  mutate(age = (age - median(age))) %>%
  glm(unarmed ~ race + age + flee + body_camera + signs_of_mental_illness, data = ., family = binomial)
glm5 <- reg_data %>%
  filter(!is.na(age)) %>%
  mutate(age = (age - median(age))) %>%
  glm(unarmed ~ race*age + race*flee + race*body_camera + race*signs_of_mental_illness, data = ., family
```

```r
stargazer(glm4,
          glm5,
          title = "Regression to Estimate the Effect of Several Factors on Probability Unarmed",
          header = F)
```

The results of this regression suggest that age still has a very strong *negative* effect on the probability the suspect was unarmed, but only for non-black victims. The officer having a body camera increases the chances that the victim is unarmed, but this only applies to non-black victims, as the interaction term is not statistically significant. However, while having a mental illness does not increase the probability of being unarmed for non-black victims, it strongly increases the risk for black victims (OR = 3.97, $p < .01$).

Table 4: Regression to Estimate the Effect of Several Factors on Probability Unarmed

	Dependent variable:	
	unarmed	
	(1)	(2)
race	0.444*	0.254
	(0.239)	(0.353)
age	−0.033***	−0.050***
	(0.010)	(0.013)
flee	0.573**	0.456
	(0.236)	(0.303)
body_camera	0.617*	0.956**
	(0.346)	(0.432)
signs_of_mental_illness	−0.101	−0.585
	(0.275)	(0.357)
race:age		0.049**
		(0.021)
race:flee		0.364
		(0.492)
race:body_camera		−0.803
		(0.731)
race:signs_of_mental_illness		1.378**
		(0.575)
Constant	−2.544***	−2.444***
	(0.188)	(0.205)
Observations	932	932
Log Likelihood	−286.505	−280.342
Akaike Inf. Crit.	585.010	580.684
Note:	*p<0.1; **p<0.05; ***p<0.01	