# Counties at risk for police killings based on racial composition and GDP

*Jeffrey Cheng, Ruida Zeng, & Arjun Bansal*

*4/7/2019*

INTRODUCTION: We analyzed a data set we found on fivethirtyeight on police killings throughout the United States and tried to find some correlations between the number of police killings and the racial composition and GDP of each location which we differentiated with their unique FIPS code. We used our results to look at the data sets of demographic data and identify which counties are at risk for police killings. This project will be useful to someone because it could help them figure out appropriate places for them to live if safety is of high priority to them. It could also be useful to city governments in evaluating how much training regarding self-defense needs to be given to their police force.

DATA: Our police killings data was obtained from fivethirtyeight and contains all unintentional killings of people by police in 2015 along with name, county, and various county statistics. The GCP_Release_1 dataset countains a list of all the counties in the US with their GDP for 4 categories: total, private goods, private services, and government. The cc-est2017-alldata dataset contains population estimates for every county with categories of total population, white male, white female, black male, black female, etc.

```
library(fivethirtyeight)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.3.1
## v readr   1.3.1       v forcats 0.3.0
```

```
## -- Conflicts ------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(choroplethr)
```

```
## Loading required package: acs
```

```
## Loading required package: XML
```

```
##
## Attaching package: 'acs'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:base':
##
##     apply
```

```
library(choroplethrMaps)
library(ggplot2)
```

```r
#load data
data("police_killings")
data("df_pop_county")
gdp_data <- read.csv("~/Desktop/Math 2820/GCP_Release_1.csv")
race_data <- read.csv("~/Desktop/Math 2820/cc-est2017-alldata.csv")


#examine data
head(police_killings)
```

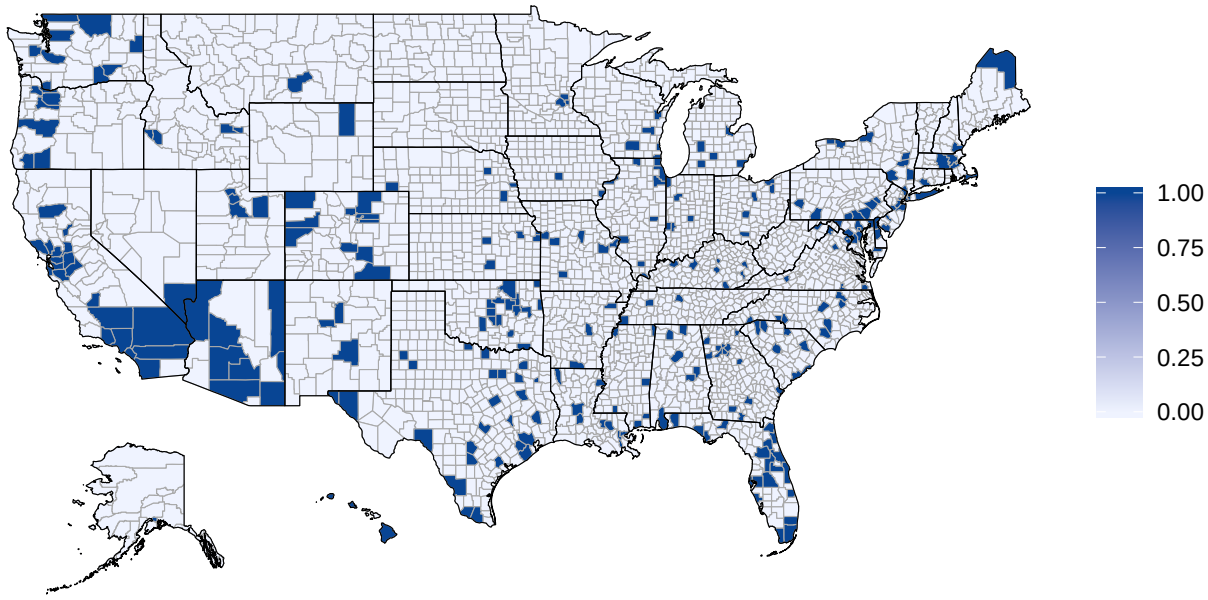
```
## # A tibble: 6 x 34
##   name    age gender raceethnicity month   day  year streetaddress city
##   <chr> <int> <chr>  <chr>         <chr> <int> <int> <chr>         <chr>
## 1 A'do~    16 Male   Black         Febr~    23  2015 Clearview Ln  Mill~
## 2 Aaro~    27 Male   White         April     2  2015 300 block Ir~ Pine~
## 3 Aaro~    26 Male   White         March    14  2015 22nd Ave and~ Keno~
## 4 Aaro~    25 Male   Hispanic/Lat~ March    11  2015 3000 Seminol~ Sout~
## 5 Adam~    29 Male   White         March    19  2015 364 Hiwood A~ Munr~
## 6 Adam~    29 Male   White         March     7  2015 18th St and ~ Phoe~
## # ... with 25 more variables: state <chr>, latitude <dbl>,
## #   longitude <dbl>, state_fp <int>, county_fp <int>, tract_ce <int>,
## #   geo_id <dbl>, county_id <int>, namelsad <chr>,
## #   lawenforcementagency <chr>, cause <chr>, armed <chr>, pop <int>,
## #   share_white <dbl>, share_black <dbl>, share_hispanic <dbl>,
## #   p_income <int>, h_income <int>, county_income <int>,
## #   comp_income <dbl>, county_bucket <int>, nat_bucket <int>, pov <dbl>,
## #   urate <dbl>, college <dbl>
```

```r
dim(police_killings)
```

```
## [1] 467  34
```

```r
#showing killings on map
kills_df <- data.frame(region = police_killings$county_id, value = 1)
county_df <- data.frame(region = df_pop_county$region, value = 0)
complete_kills_df <- rbind(kills_df, county_df)
complete_kills_df <- distinct(complete_kills_df, region, .keep_all = TRUE)
county_choropleth(complete_kills_df,
                  title  = "Locations of killings by US police in 2015", num_colors = 1)
```

## Locations of killings by US police in 2015

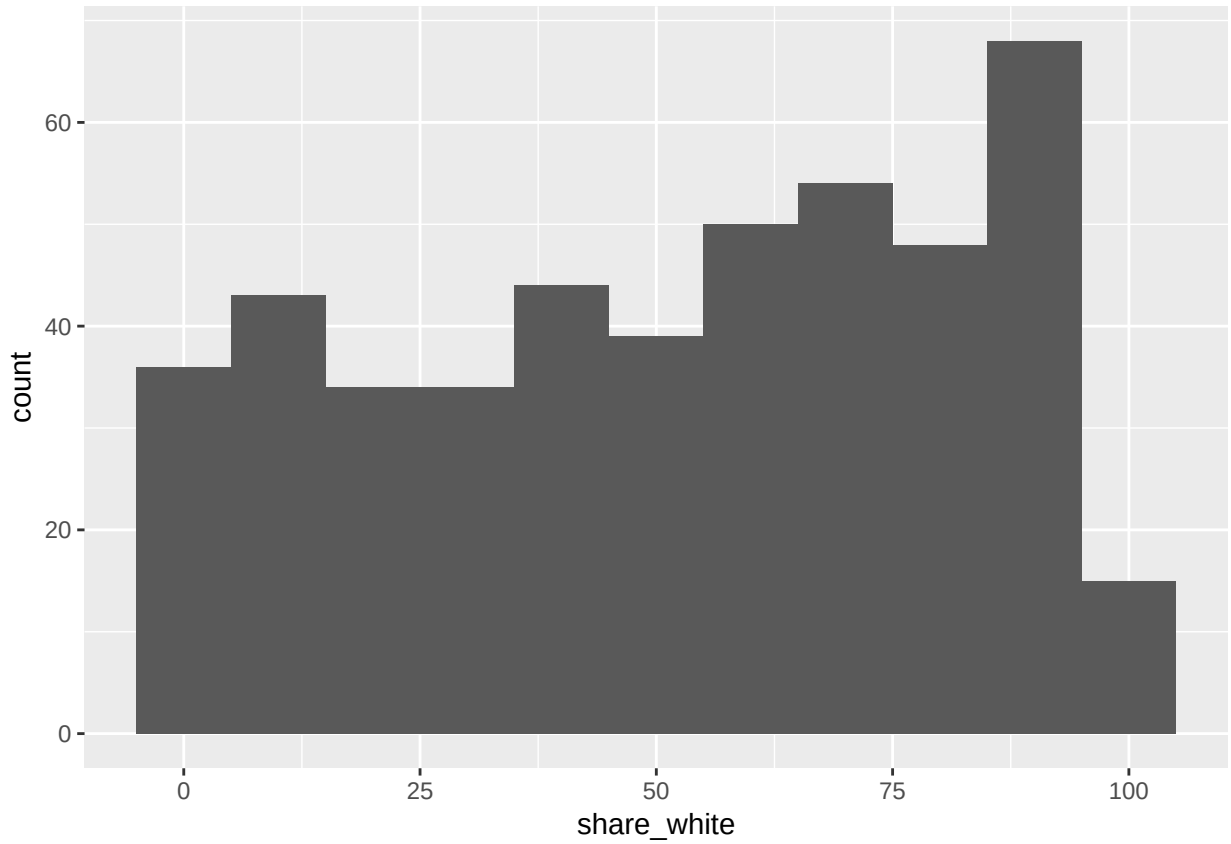



There are only 467 killings for the year 2015. They are in very high density in Southern California and New Mexico. Other than that, the highest concentration is in counties in the South and Southeast of the US.

```
#visualizing data
ggplot(police_killings, aes(x=share_white)) + geom_bar(binwidth = 10)
```

```
## Warning: `geom_bar()` no longer has a `binwidth` parameter. Please use
## `geom_histogram()` instead.
```

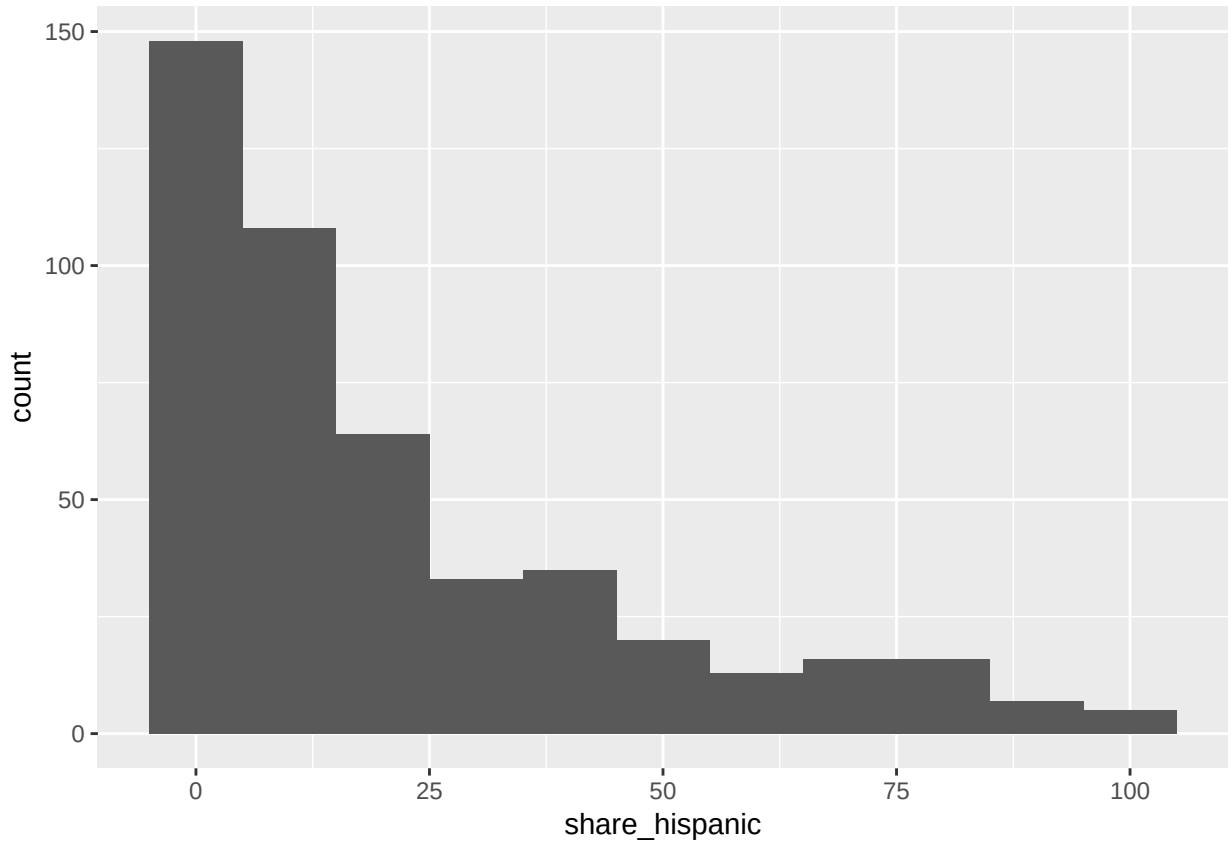
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```




```
ggplot(police_killings, aes(x=share_hispanic)) + geom_bar(binwidth = 10)
```

```
## Warning: `geom_bar()` no longer has a `binwidth` parameter. Please use
## `geom_histogram()` instead.
```

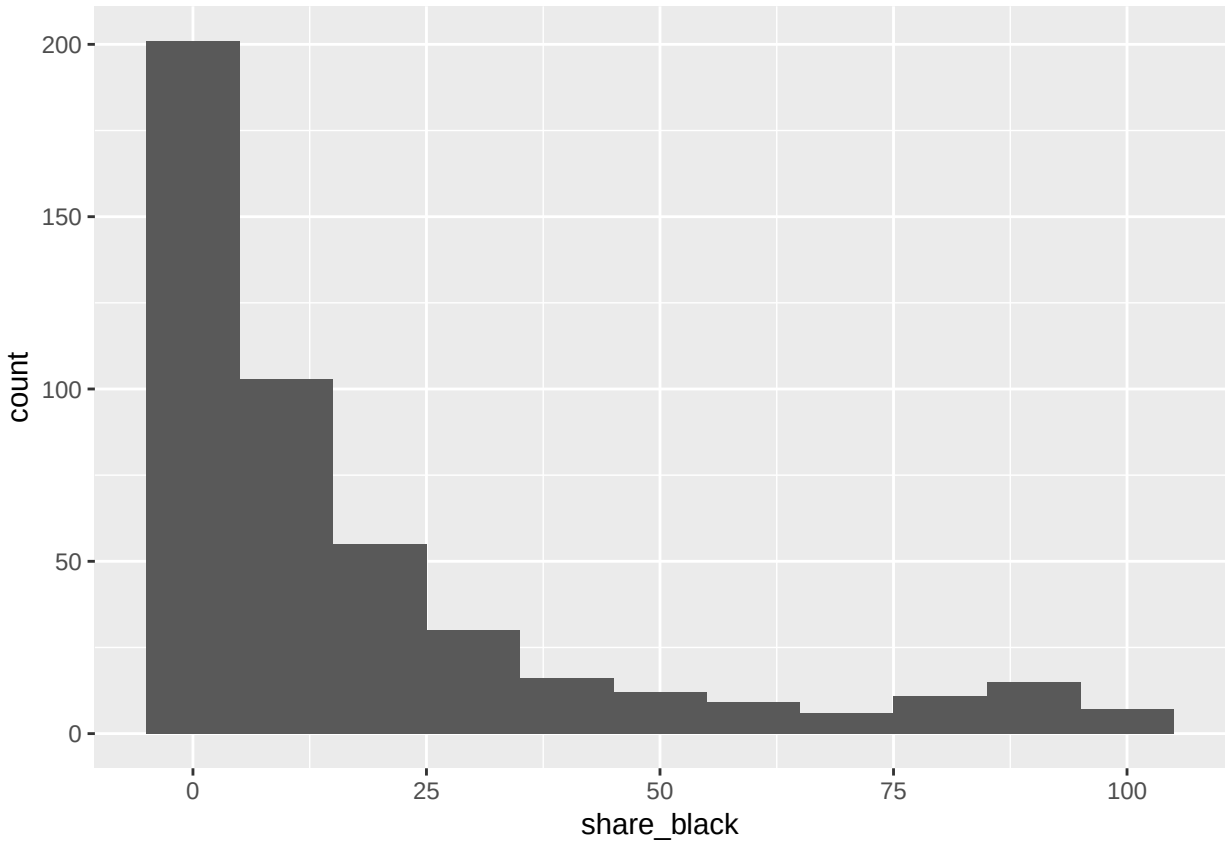
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```




```
ggplot(police_killings, aes(x=share_black)) + geom_bar(binwidth = 10)
```

```
## Warning: `geom_bar()` no longer has a `binwidth` parameter. Please use
## `geom_histogram()` instead.

## Warning: Removed 2 rows containing non-finite values (stat_bin).
```




It's clear that higher white populations are correlated with killings up to a critical point of around 90-100%. Also, populations with low populations of black and hispanic people have more killings. The curve appears to be exponential for the black and hispanic populations.

```
#mutating data sets

#adding a kill variable
police_killings <- police_killings %>%
  mutate(kill = 1)

#filtering to the year we want
race_data <- race_data %>%
  filter(YEAR == 5, AGEGRP == 3)

#In the BEA data, populations are given as raw numbers so we had to calculate proportions on our own
race_data <- race_data %>%
  mutate(kill = 0, county_fp = STATE * 1000 + COUNTY, share_white = ((WA_MALE + WA_FEMALE) / TOT_POP) *


#joining data sets to model race data

#created two data frames with the data we want
police_df <- data.frame(county_fp = police_killings$county_id, share_white = police_killings$share_whit
race_df <- data.frame(county_fp = race_data$county_fp, share_white = race_data$share_white, share_black

#appends both data frames together on top of each other
all_race_data <- rbind(police_df, race_df)
```

```r
#removes duplicate counties
all_race_data <- distinct(all_race_data, county_fp, .keep_all = TRUE)


#race models
white_model <- lm(kill ~ share_white, all_race_data)
summary(white_model)
```

```
##
## Call:
## lm(formula = kill ~ share_white, data = all_race_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47848 -0.09671 -0.03904 -0.02162  0.99504
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4861107  0.0195048   24.92   <2e-16 ***
## share_white -0.0048308  0.0002333  -20.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2754 on 3138 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.1202, Adjusted R-squared:   0.12
## F-statistic: 428.9 on 1 and 3138 DF,  p-value: < 2.2e-16
```

```r
black_model <- lm(kill ~ share_black, all_race_data)
summary(black_model)
```

```
##
## Call:
## lm(formula = kill ~ share_black, data = all_race_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27551 -0.08962 -0.07528 -0.07330  0.92791
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0720900  0.0060259  11.963  < 2e-16 ***
## share_black 0.0022252  0.0002941   7.566 5.01e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.291 on 3138 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.01792,    Adjusted R-squared:  0.0176
## F-statistic: 57.25 on 1 and 3138 DF,  p-value: 5.013e-14
```

```r
hispanic_model <- lm(kill ~ share_hispanic, all_race_data)
summary(hispanic_model)
```

```
##
```

```
## Call:
## lm(formula = kill ~ share_hispanic, data = all_race_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27502 -0.09221 -0.07990 -0.07547  0.92885
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.0711544  0.0063654  11.178   <2e-16 ***
## share_hispanic 0.0020860  0.0003177   6.567    6e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2916 on 3138 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.01355,    Adjusted R-squared:  0.01324
## F-statistic: 43.12 on 1 and 3138 DF,  p-value: 6.003e-11
```

White model: R^2 = 0.12, has a slightly negative slope. This is likely due to the anomaly at the very right end of the data when share of white population is 90-100%, the amount of killings is very low.

Black model: R^2 = 0.01, has a slightly positive slope. This low R^2 value is because the data is not linear and we cannot use black proportion as a predictor for killings.

Hispanic model: R^2 = 0.01 as well, has a slightly positive slope. This low R^2 value is because the data is not linear and we cannot use black proportion as a predictor for killings.

From these three, if we were to use one race population as a predictor for killings, white would be the best one.

```
confint(white_model)
```

```
##                    2.5 %       97.5 %
## (Intercept)  0.447867249  0.524354138
## share_white -0.005288201 -0.004373432
```

```
confint(black_model)
```

```
##                   2.5 %      97.5 %
## (Intercept) 0.060275012 0.083905040
## share_black 0.001648553 0.002801801
```

```
confint(hispanic_model)
```

```
##                     2.5 %      97.5 %
## (Intercept)    0.058673672 0.083635098
## share_hispanic 0.001463167 0.002708916
```

The 95% confidence interval for our white model has all negative values, so we can see that by this model, increasing white population has negative effect on killings. In a linear model, this is likely due to the huge drop in killings once white population because 90-100% of the population. We can see that if a linear model is used, increasing white population leads to less killings.

The 95% confidence interval for our black and hispanic models have all positive values, so we can see that by this model, increasing black and hispanic populations have positive effect on killings. From this, we can see that if a linear model is used, increasing black and hispanic populations leads to more killings.

Overally, all 3 race models are not very useful to us.

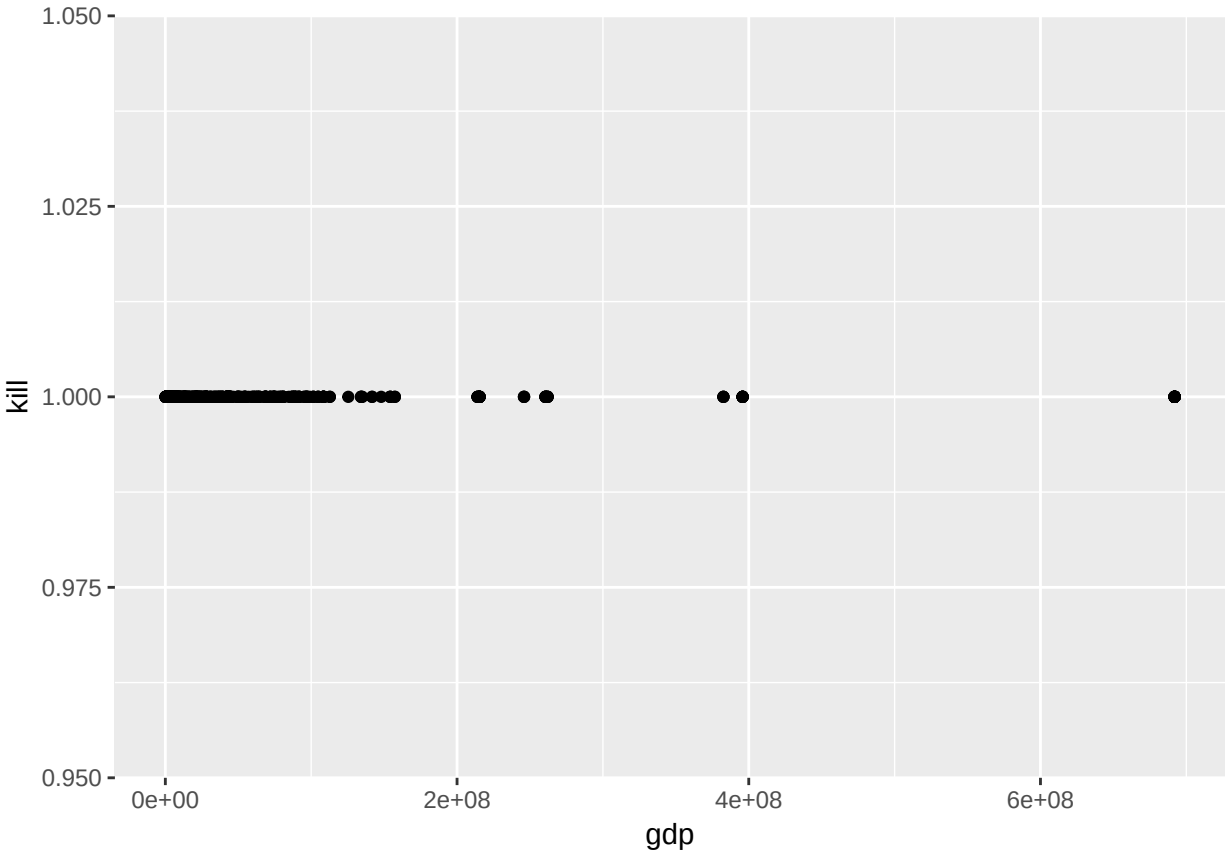
```r
#mutating data sets to model economic data
gdp_data <- gdp_data %>%
  filter(LineData==1)
police_econ_df <- data.frame(county_fp = police_killings$county_id, kill = 1)
gdp_econ_df <- data.frame(county_fp = gdp_data$FIPS, kill = 0, gdp = gdp_data$gdp)
all_gdp_data <- full_join(police_econ_df, gdp_econ_df, by = "county_fp")
head(all_gdp_data)
```

```
##   county_fp kill.x kill.y       gdp
## 1      1051      1      0   1639473
## 2     22079      1      0   5204707
## 3     55059      1      0   6193261
## 4      6037      1      0 691948578
## 5     39153      1      0  28494352
## 6      4013      1      0 215381372
```

```r
all_gdp_data <- all_gdp_data %>%
  mutate(kill = (kill.x + kill.y))

#plot of gdp vs kills
ggplot(all_gdp_data, aes(x=gdp, y = kill)) + geom_point()
```

```
## Warning: Removed 2819 rows containing missing values (geom_point).
```




Its clear to see that most of the kills are in counties with lower GDP. Specifically, counties with a GDP of less than 150,000,000 are at much higher risk.

```r
#economic models
gdp_model <- lm(kill ~ gdp, all_gdp_data)
summary(gdp_model)
```

```
##
## Call:
## lm(formula = kill ~ gdp, data = all_gdp_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.640e-16 -6.560e-16 -6.160e-16 -5.180e-16  2.385e-13
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.000e+00  5.868e-16  1.704e+15   <2e-16 ***
## gdp         -1.812e-24  3.411e-24 -5.310e-01    0.595
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111e-14 on 462 degrees of freedom
##   (2819 observations deleted due to missingness)
## Multiple R-squared:    0.5,  Adjusted R-squared:  0.4989
## F-statistic:   462 on 1 and 462 DF,  p-value: < 2.2e-16
```

```r
confint(gdp_model)
```

```
##                     2.5 %       97.5 %
## (Intercept)  1.000000e+00 1.000000e+00
## gdp         -8.515994e-24 4.891043e-24
```

This model works decently well: much better than all the race models due to its R^2 value of 0.5. However, the 95% confidence interval has both positive and negative values, so we cannot be sure that gdp has a negative relationship with killings.

Because we are still curious about having a combined model, we will try a model using both share white and GDP.

```r
#combining variables
all_data <- full_join(all_gdp_data, all_race_data, by = "county_fp")
head(all_data)
```

```
##   county_fp kill.x kill.y       gdp kill.x.x share_white share_black
## 1      1051      1      0   1639473        1        60.5        30.5
## 2     22079      1      0   5204707        1        53.8        36.2
## 3     55059      1      0   6193261        1        73.8         7.7
## 4      6037      1      0 691948578        1         1.2         0.6
## 5     39153      1      0  28494352        1        92.5         1.4
## 6      4013      1      0 215381372        1         7.0         7.7
##   share_hispanic kill.y.y
## 1            5.6        1
## 2            0.5        1
## 3           16.8        1
## 4           98.8        1
## 5            1.7        1
## 6           79.0        1
```

```r
all_data <- all_data %>%
  mutate(kill = (kill.x + kill.y))


#combined model
combined_model <- lm(kill + share_white ~ gdp, all_data)
summary(combined_model)
```

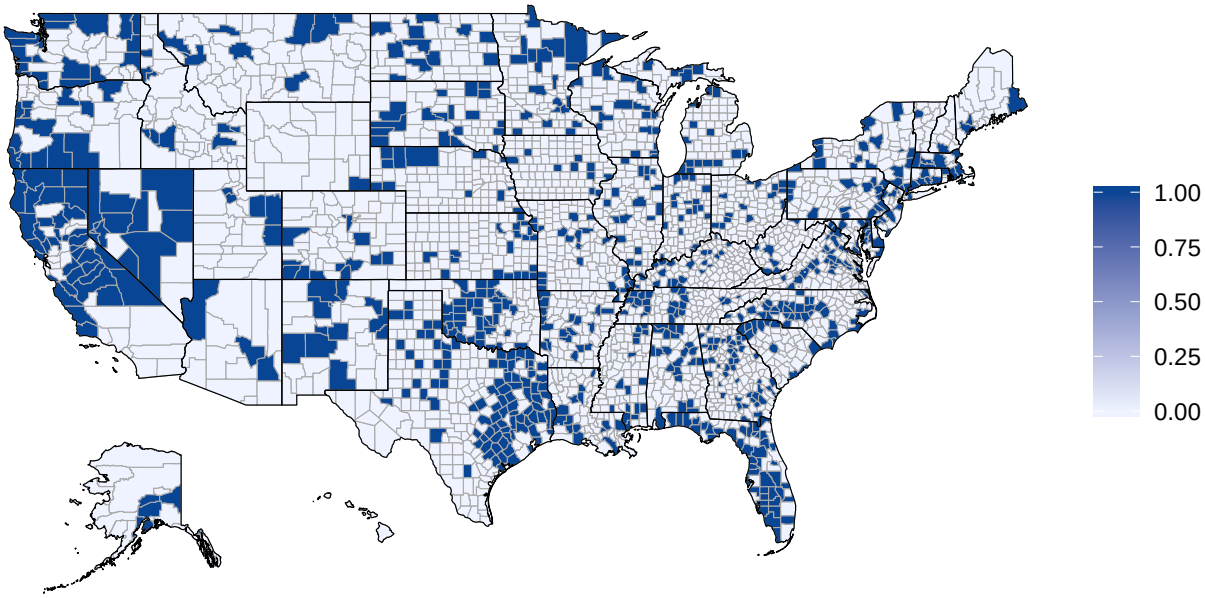
```
##
## Call:
## lm(formula = kill + share_white ~ gdp, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.962 -18.536   2.525  20.725  48.438
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.915e+01  1.418e+00   41.72   <2e-16 ***
## gdp         -1.046e-07  8.222e-09  -12.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.78 on 459 degrees of freedom
##   (2872 observations deleted due to missingness)
## Multiple R-squared:  0.2608, Adjusted R-squared:  0.2592
## F-statistic: 161.9 on 1 and 459 DF,  p-value: < 2.2e-16
```

This model is better than all the race models, but still not as good as only GDP because of the problem with kill ~ white not actually being linear. We will not do a confidence interval for this model because of potential correlation between the variables.

If we want to do a linear model, the best way to predict killings by race would be to go off share white because of its R^2 value of 0.12 and definitively negative correlation with killings from our 95% confidence interval. However, this is clearly wrong: clearly increase in white proportion leads to increasing numbers of kills until white proportion is 90-100% of the population so our data is not well modeled by a linear model. The best model to use would likely be the GDP model to predict killings in specific counties.

```r
#showing at risk counties on map
at_risk <- all_data %>%
  filter(share_white > 70, share_white < 90, gdp < 150000000)
atrisk_df <- data.frame(region = at_risk$county_fp, value = 1)
counties_df <- data.frame(region = df_pop_county$region, value = 0)
complete_risk_df <- rbind(atrisk_df, county_df)
complete_risk_df <- distinct(complete_risk_df, region, .keep_all = TRUE)
county_choropleth(complete_risk_df,
                  title  = "Locations of at risk counties for killings", num_colors = 1)
```

## Locations of at risk counties for killings




In conclusion, the at risk counties are generally in poorer areas with high, but not all, white populations. We think more concrete anlyses could have been attempted on a data set with more kills because the number of kills compared to the number of counties is extremely small. Perhaps a future experiment could aggregate killing data from multiple years to get more data points. We think that the killings go up as white population increases because of racism and historical animosity against communities of color, but decreases sharply as white populations get up to 90-100% because those populations are predominantly white and would be less vulnerable to racist behavior. This also makes sense with the GDP data: less wealthy counties would typically be more conservative.