# Counties at Risk for Police Killings Based on Racial Composition and GDP

Jeffrey Cheng, Ruida Zeng, Arjun Bansal

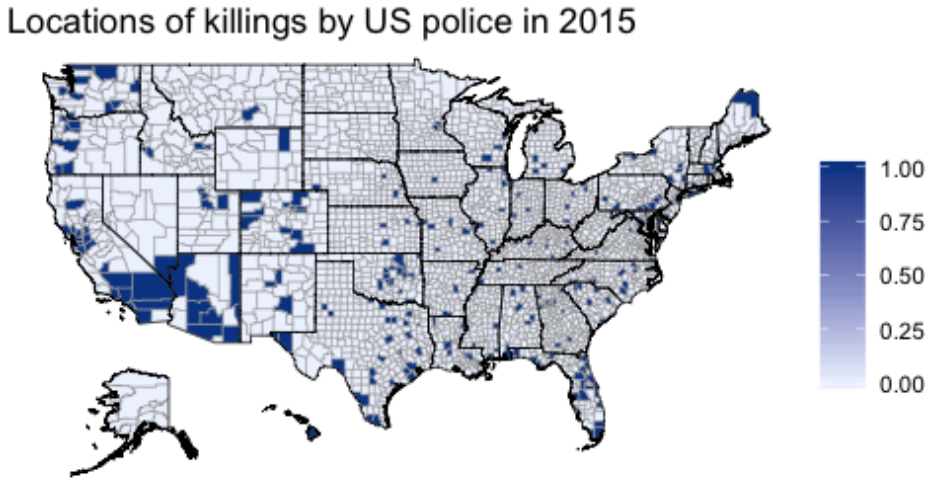
04-07-19

# 1 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.

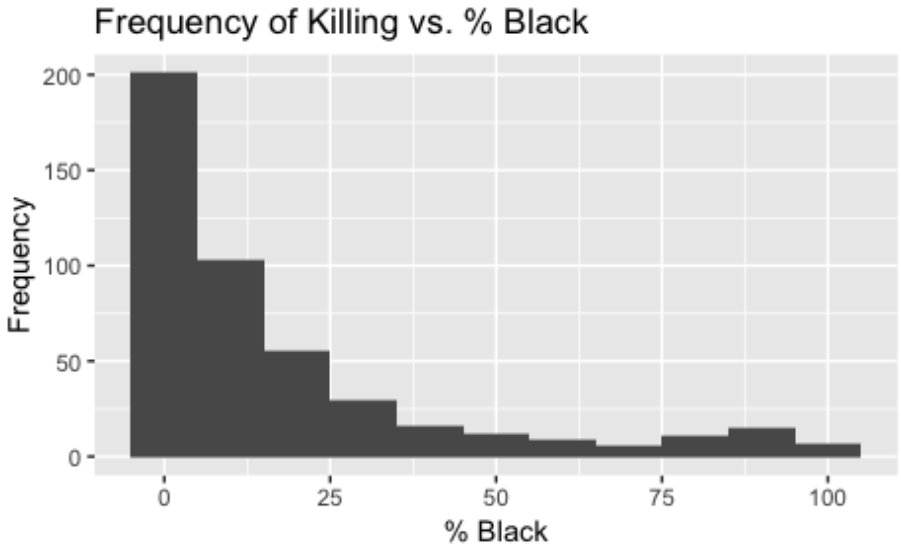
Three datasets were used: "Police Killings", "GCP Release", and "cc-est2017-alldata" Our police killings data was obtained from fivethirtyeight and contains all unintentional killings of people by police in 2015 along with name, county, and various county statistics. The GCP Release dataset countains 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.

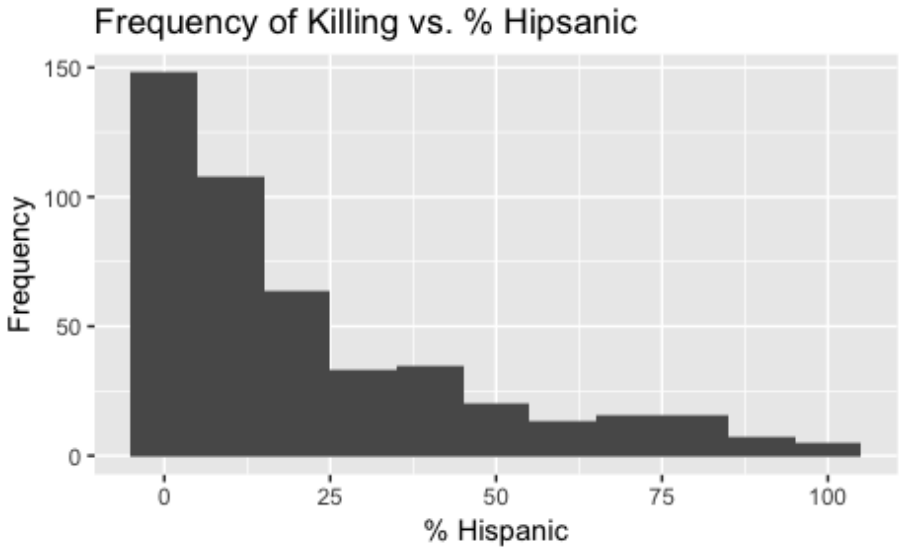
# 2 Examining Data

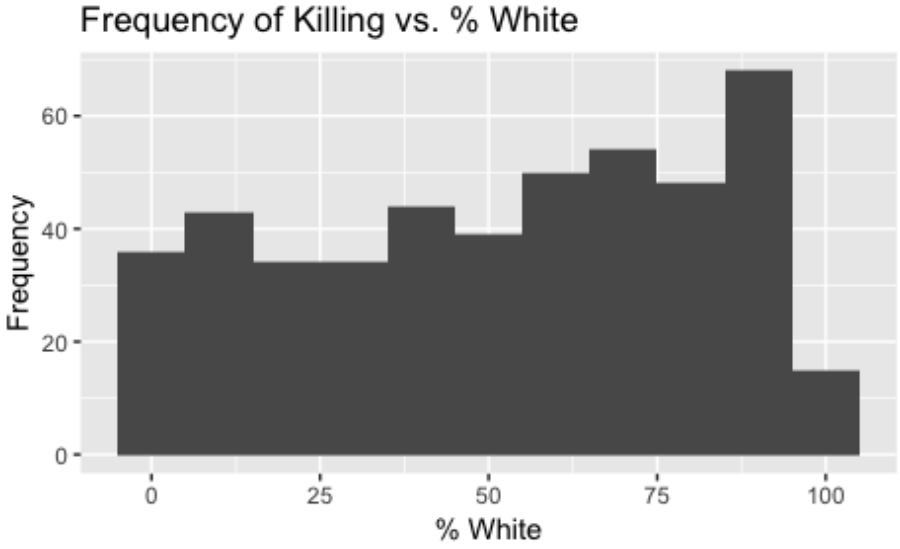



There are only 467 killings for the year 2015. They are in very high density in Southern California and New Mexico. Other than that, the highest concentra-

tion is in counties in the South and Southeast of the US.






It's clear that higher white populations are correlated with killings up to a critical point of around 90-100 percent. 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.

# 3 Testing Correlation: Race $\iff$ Killings

## 3.1 Linear Regression Models

First, we must combine the racial data for all the US counties from "cc-est2017-alldata" to the "police killings" dataset.

```
#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 "cc-est2017-alldata" dataset, 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) * 100, share_black = ((BA_
      MALE + BA_FEMALE) / TOT_POP) * 100, share_hispanic = ((H_MALE
      + H_FEMALE) / TOT_POP) * 100)

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

```
    _killings$share_black, share_hispanic = police_killings$share_
    hispanic, kill = police_killings$kill)
race_df <- data.frame(county_fp = race_data$county_fp, share_white
    = race_data$share_white, share_black = race_data$share_black,
    share_hispanic = race_data$share_hispanic, kill = race_data$
    kill)

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

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

Next, we will create linear regression models for each race vs. number of police killings.

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

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

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

Results from the linear regression models:

```

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

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


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

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


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

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 percent, 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.

## 3.2 Confidence Intervals

We will create confidence intervals for the slopes of these three linear regression models as well.

```
                    2.5 %       97.5 %
(Intercept)  0.447867249  0.524354138
share_white -0.005288201 -0.004373432
                    2.5 %       97.5 %
(Intercept) 0.060275012 0.083905040
share_black 0.001648553 0.002801801
                     2.5 %       97.5 %
(Intercept)     0.058673672 0.083635098
share_hispanic 0.001463167 0.002708916
```

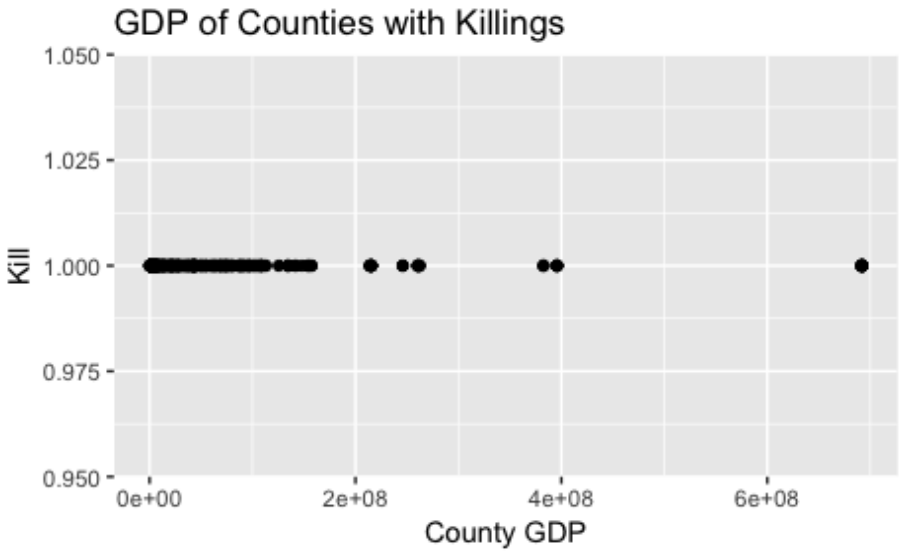
The 95 percent confidence interval for our white model has all negative values, so we can see that by this model, increasing white population has negative effect on killings. In a linear model, this is likely due to the huge drop in killings once white population because 90-100 percent of the population.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

The 95 percent 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.

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

# 4 Testing Correlation: GDP $\iff$ Killings

## 4.1 Overview of data




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.

## 4.2 Linear Regression Model

We combined the "GCP Release" data with the "Police Killings" data into a format that we can perform linear regression analysis on. Then, we created a linear regression model for GDP vs. Killings.

```
#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)
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() + labs(x
    = "County GDP", y = "Kill") + ggtitle("GDP of Counties with
    Killings")

#economic model
gdp_model <- lm(kill ~ gdp, all_gdp_data)
summary(gdp_model)
confint(gdp_model)
```

These are the results from the 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

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
```

### 4.3 Confidence Interval

This model works decently well: much better than all the race models due to its $R^2$ value of 0.5.

Next, we created a confidence interval for the slope of the linear model.

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

The 95 percent confidence interval has both positive and negative values, so we cannot be sure that gdp has a negative relationship with killings.

## 5 Testing Multivariate Correlation: GDP and Race $\iff$ Killings

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

```
#combining variables
all_data <- full_join(all_gdp_data, all_race_data, by = "county_fp"
    )
head(all_data)
all_data <- all_data %>%
  mutate(kill = (kill.x + kill.y))


```

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

These are the results from the combined model:

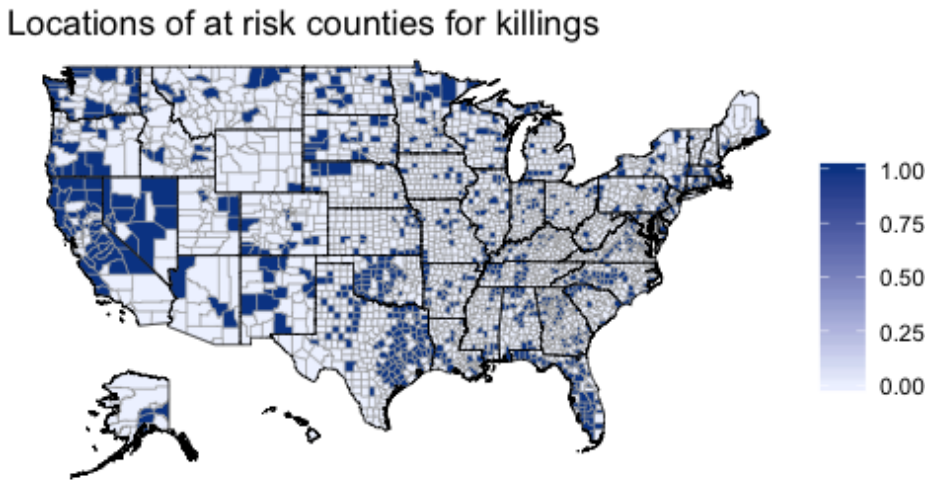
```

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

Residual standard error: 26.78 on 459 dgrees 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 percent confidence interval. However, this is clearly wrong: clearly increase in white proportion leads to increasing numbers of kills until white proportion is 90-100 percent 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.

# 6 Conclusion: Counties at Risk for Police Killings

We defined at risk counties as counties with a GDP less than 150000000 and a white population percentage between 70 and 90 percent.




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 percent 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.