Noah's Data Analysis

Noah Chasek-Macfoy

```
library(dplyr)
library(tidyr)
library(stringr)
library(lme4)
library(zoo)
library(ggplot2)
```

Regressions

Load Data

```
file <- "./Data/PDI (TEST DATA ONLY)_Use_of_Force_Cincinnati_v2_ANNOTATED.csv"
data <- read.csv(file, stringsAsFactors = FALSE)
```

Prepare Data

Note there are a number of census tracts in the data outside of Ohio (code = 39) and outside of Hamilton county (code = 061). I will leave them for now.

County Codes

```
as.data.frame(table(data$COUNTY_FIPS[-1]))
```

```
##                    Var1  Freq
## 1                   015     3
## 2                   017     2
## 3                   025     3
## 4                   037     2
## 5                   047     9
## 6                   051     1
## 7                   061 20967
## 8                   075     1
## 9                   099     3
## 10                  113     1
## 11                  117     9
## 12                  121     1
## 13 CENSUS_GEO_API_ERROR   255
## 14      MISSING_LAT_LNG   369
```

State Codes

```
as.data.frame(table(data$STATE_FIPS[-1]))
```

```
##                   Var1  Freq
## 1                   06     1
## 2                   12     3
## 3                   13     1
## 4                   21    14
## 5                   36     9
```

```
## 6                   39 20973
## 7                   45     1
## 8 CENSUS_GEO_API_ERROR   255
## 9      MISSING_LAT_LNG   369
```

```
# combine state, county, and census tract ids
tract_short <- paste0(data$STATE_FIPS, data$COUNTY_FIPS, data$LOCATION_CENSUS_TRACT)

cols <- c("SUBJECT_RACE_CLEAN", "INCIDENT_DATE_CLEAN", "TYPE_OF_FORCE_USED_CLEAN")
data <- cbind(data[cols], tract_short, stringsAsFactors=FALSE)[-1,]
names(data)[1:3] <- c("race", "date", "force")
```

Look at non-numeric tract numbers

```
mask <- is.na(as.numeric(tract_short))
```

```
## Warning: NAs introduced by coercion
```

```
sum(mask) # 625
```

```
## [1] 625
```

```
unique(tract_short[mask])
```

```
## [1] "cleaned versioncleaned versioncleaned version"
## [2] "MISSING_LAT_LNGMISSING_LAT_LNGMISSING_LAT_LNG"
## [3] "CENSUS_GEO_API_ERRORCENSUS_GEO_API_ERRORCENSUS_GEO_API_ERROR"
```

Remove non-numeric census tract rows

```
mask <- !is.na(as.numeric(tract_short))
```

```
## Warning: NAs introduced by coercion
```

```
data <- data[mask,]
data$tract_short <- as.numeric(data$tract_short)
```

```
## Warning: NAs introduced by coercion
```

```
# save copy for visualization
viz_data <- data
```

Note: only Black and White are over 5%

```
# view racial makeup of force incidents
df <- as.data.frame(table(data$race), stringsAsFactors = FALSE)
df$pct <- (df$Freq/length(data$race))*100
df
```

```
##              Var1  Freq          pct
## 1           ASIAN    21  0.099990477
## 2           BLACK 15095 71.874107228
## 3        HISPANIC    95  0.452337873
## 4         MISSING     1  0.004761451
## 5    MULTI_RACIAL    12  0.057137415
## 6      OTHER_RACE    59  0.280925626
## 7 VALUE_NOT_KNOWN   368  1.752214075
## 8           WHITE  5350 25.473764403
```

```
# remove races less than 5% of incidence
races <- df[df$pct > 5, 1]
```

```
mask <- data$race %in% races
data <- data[mask,]
data$race <- tolower(data$race)
```

Sources of measurment and selection bias in the data:
Excluding small racial groups does not affect the number of force incidents per person ratio within non-excluded racial groups. Small racial groups must be excluded because even a small amount of race missclassification between races with disparate rates of victimization would greatly skew the number of force incidents per person in smaller groups. Also note that large bias in which race is most frequently selected into "unknown" has the potential to bias estimates of uses of force by race. This said 'unknown' accounts for 1.8% of the data and thus will likely have little impact on the estimates of racial groups which make up a greater than 5% share of the incidents.

Load census tract population by race data.

```
file <- "./Data/ACS_5YR_racial_population_demographics.csv"
pop <- read.csv(file, stringsAsFactors = FALSE)

# select relevant cols
cols <- c("tract_short", "total_population", "white","black_or_african_american","other","total_hispani
          ,"asian" )
pop <- pop[cols]
names(pop)[c(2,4,5,6)] <- c("total_pop", "black", "other_race", "hispanic")
pop$pct_black <- (pop$black/pop$total_pop) * 100

# save copy for visualizations
viz_pop <- pop
```

Note: The number of census tracts in the population data and the use of force data do not match up.

183 unique census tracts in population by race data.

```
## Remove census tracts in force incidents not found in population data
length(unique(pop$tract_short))
```

```
## [1] 183
```

195 unique census tracts in use of force data.

```
length(unique(data$tract_short))
```

```
## [1] 195
```

We see that the force incidents census tracts are a super set of the population by race data census tracks. That means there are no census tracks that we will analyze where no use of force incident was recorded between 1996 and 2018.

```
setdiff(pop$tract_short, data$tract_short)
```

```
## numeric(0)
```

```
# Remove census tracts in force incidents not found in population data
mask <- data$tract_short %in% pop$tract_short
count_data <- data[mask,]
```

```
## Get use of force counts by census tract by race
count_data <- group_by(count_data, race, tract_short) %>%
  summarise(count = n())

# transform to by tract by race row format
```

```
pop <- gather(pop, race, pop_race, white, black)

## Merge population data onto count data
count_data <- left_join(pop, count_data, by=c("tract_short", "race"))
# set race tract pairs with no force incidents to count = 0
count_data$count[is.na(count_data$count)] <- 0
```

Note: There are some etreme outliers in the use of force data. For example, one tract has a max value more than 6 times greater than the 75th percentile value.

```
summary(count_data)
```

```
##   tract_short          total_pop      other_race        hispanic
##  Min.   :3.902e+10   Min.   : 780   Min.   :  0.00   Min.   :   0.00
##  1st Qu.:3.906e+10   1st Qu.:2095   1st Qu.:  0.00   1st Qu.:  21.25
##  Median :3.906e+10   Median :3279   Median :  8.00   Median :  66.00
##  Mean   :3.906e+10   Mean   :3552   Mean   : 31.98   Mean   : 107.34
##  3rd Qu.:3.906e+10   3rd Qu.:4836   3rd Qu.: 33.75   3rd Qu.: 152.25
##  Max.   :3.911e+10   Max.   :8344   Max.   :397.00   Max.   :1162.00
##      asian          pct_black           race              pop_race
##  Min.   :  0.00   Min.   : 0.01198   Length:366         Min.   :   1.0
##  1st Qu.:  5.00   1st Qu.: 6.61705   Class :character   1st Qu.: 405.8
##  Median : 26.00   Median :25.04810   Mode  :character   Median :1148.5
##  Mean   : 81.27   Mean   :32.13457                      Mean   :1664.0
##  3rd Qu.: 92.75   3rd Qu.:53.48184                      3rd Qu.:2586.2
##  Max.   :902.00   Max.   :95.10846                      Max.   :8293.0
##      count
##  Min.   :  0.00
##  1st Qu.:  2.00
##  Median : 12.00
##  Mean   : 54.64
##  3rd Qu.: 60.50
##  Max.   :658.00
```

Add all-race poverty rate by census tract.

```
file = "./Data/ACS_5YR_poverty.csv"
pov <- read.csv(file, stringsAsFactors = FALSE)

cols <- c("GEO.id2", "HC03_EST_VC01")
pov <- pov[-1 ,cols]
names(pov) <- c("tract_short", "pct_pov")
pov <- as.data.frame(lapply(pov, as.numeric)) # char to num
```

Note: there are 5 census tracts in the force count data that are not found in the poverty data, all of which outside of hamilton county however.

```
setdiff(count_data$tract_short, pov$tract_short)
```

```
## [1] 39017011123 39025041305 39025041404 39025041501 39113165100
```

There also appear to be a number of hamilton county census tracts in the poverty dataset but not in the use of force/population by race data.

```
setdiff(pov$tract_short,count_data$tract_short)
```

```
##  [1] 39061020401 39061020403 39061020404 39061020501 39061020502
##  [6] 39061020602 39061020707 39061020741 39061021001 39061021002
```

```
## [11] 39061021003 39061021101 39061021102 39061021201 39061021422
## [16] 39061021501 39061021504 39061021505 39061021506 39061021508
## [21] 39061021571 39061021602 39061021603 39061021604 39061021701
## [26] 39061022000 39061022101 39061022301 39061022302 39061023002
## [31] 39061023210 39061023501 39061023600 39061023701 39061023702
## [36] 39061024002 39061024301 39061024303 39061024322 39061024901
## [41] 39061026001 39061026102 39061026200 39061027300
```

```
# Merge with count data
count_data <- left_join(count_data, pov, by="tract_short")
```

Take a look at the observations we are loosing by exluding the tracts with no poverty data. A number of largely white census tracts with relatively low uses of force. The distribution suggest they will bias the results.

```
count_data[is.na(count_data$pct_pov),]
```

```
##     tract_short total_pop other_race hispanic asian  pct_black  race
## 1   39017011123      4902        397     1162   163 31.5177479 white
## 2   39025041305      5178          0       55   114  2.4719969 white
## 3   39025041404      4913          0       98   118  0.9159373 white
## 4   39025041501      6667         17       71    74  6.5846708 white
## 183 39113165100      2351          0        0    10 95.1084645 white
## 184 39017011123      4902        397     1162   163 31.5177479 black
## 185 39025041305      5178          0       55   114  2.4719969 black
## 186 39025041404      4913          0       98   118  0.9159373 black
## 187 39025041501      6667         17       71    74  6.5846708 black
## 366 39113165100      2351          0        0    10 95.1084645 black
##     pop_race count pct_pov
## 1       2676     0      NA
## 2       4810     1      NA
## 3       4676     1      NA
## 4       6124     1      NA
## 183       72     0      NA
## 184     1545     2      NA
## 185      128     0      NA
## 186       45     0      NA
## 187      439     0      NA
## 366     2236     1      NA
```

```
#(df$count/df$pop_race)[is.na(df$pct_pov)]
```

I consider filling the missing data with the pct_pov mean, but since the tracts seem relatively unrepresentatively low in terms of use of force the overall mean doesn't seem appropriate. I consider filling the NA values with the mean poverty measure for geographically proximate tracts or that of tracts that fall within the same range of uses of force per person. But I will leave that judgement for future work and exclude these observations here.

```
count_data <- count_data[!is.na(count_data$pct_pov),]
```

Add part 1 crimes per person per census tract to data.

```
file <- "./Data/PDI (TEST DATA ONLY)_Crime_Incidents_Cincinnati_2014-2015_ANNOTATED.csv"
crime <- read.csv(file, stringsAsFactors = FALSE)

s_fips <- str_pad(crime$STATE_FIPS, 2, "left", "0")
c_fips <- str_pad(crime$COUNTY_FIPS, 3, "left", "0")
tract <- str_pad(crime$LOCATION_CENSUS_TRACT, 6, "left", "0")
```

```
tract_short <- paste0(s_fips, c_fips , tract)

cols <- c("UNKNOWN_FIELD_TYPE.26")
crime <- cbind(crime[cols], tract_short, stringsAsFactors=FALSE)[-1,]
names(crime)[1] <- "ucr_type"

# Remove string valued census tract rows
mask <- !is.na(as.numeric(crime$tract_short))
```

```
## Warning: NAs introduced by coercion
```

```
crime <- crime[mask,]
crime$tract_short <- as.numeric(crime$tract_short)
```

I am unsure of the UCR classification of "UNAUTHORIZED USE", but will include it in the part 1 count. I checked the UCR handbook on this issue and did not find a direct indication of what "UNAUTHORIZED USE" might refer to, but it is listed as 701 in "UNKNOWN_FIELD_TYPE.4" aka "UCR" which corresponds to the listing of "moto vehicle theft" in the handbook as category 7. Other crime codes in "UNKNOWN_FIELD_TYPE.4" correspond to the crime categories in the handbook.

Note: Not counting crimes whose UCR group is not stated. I am assuming these were not on among UCR mandatory reporting crimes.

```
crime$part1 <- 0
crime$part1[!(crime$ucr_type %in% c("", "PART 2 MINOR"))] <- 1

crime$murder <- 0
crime$murder[crime$ucr_type == "HOMICIDE"] <- 1

# get counts part1 crimes per tract
by_tract <- group_by(crime, tract_short) %>% summarise(part1_count = sum(part1), murder_count = sum(mur
crime <- by_tract

# merge with count data
count_data <- left_join(count_data, crime, by= "tract_short")
```

Note: there are 8 tracts in the crime data note found in the use of force data set.

```
setdiff(crime$tract_short, data$tract_short)
```

```
## [1]  8031004107 18029080202 21015070401 39061020401 39061021201 39061021571
## [7] 39061022301 39061024301 39165030700
```

```
#length(setdiff(count_data$tract_short,crime$tract_short))
```

As with the poverty data I could choose to fill in values for these tracts without crime data but I will delete them.

```
count_data <- count_data[!is.na(count_data$part1_count),]

# number of part1 crimes per 100 people
count_data$part1_per_capita <- (count_data$part1_count/count_data$total_pop) * 100

count_data$murder_per_capita <- (count_data$murder_count/count_data$total_pop) * 1000
```

Note: that outlier with 81 part 1 crimes per 100 people... Seems suspicious.

```r
summary(count_data$part1_per_capita)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.565   9.298  11.600  15.550  81.280
```

Modeling

Specification 1: Race and census tract

$$y = n_j e^{\beta_0 + \beta_{black} X_{black} + \alpha_j}$$

Where n_j is the population of out each racial group in each census tract, and X_{black} is an indicator of black race, α_j is the partial-pooling (i.e. random effects) intercept deviation for each race per census tract group, and y is the number of use of force indicents against a racial group in a county.

```r
# set factor with level order such that white will be reference group
count_data$race <- factor(count_data$race, levels=c("white", "black"))
count_data$tract_short <- factor(count_data$tract_short)
```

Fit Model

```r
m1 <- glmer(count ~ (1| tract_short) + race , offset=log(pop_race), data=count_data, family = poisson())

summary(m1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: count ~ (1 | tract_short) + race
##    Data: count_data
##  Offset: log(pop_race)
##
##      AIC      BIC   logLik deviance df.resid
##   5704.7   5716.0  -2849.4   5698.7      311
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -11.347  -1.326  -0.175   1.931  39.107
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  tract_short (Intercept) 3.274    1.809
## Number of obs: 314, groups:  tract_short, 157
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.76240    0.14658  -32.49   <2e-16 ***
## raceblack    1.38849    0.01926   72.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## raceblack -0.075
```

Interpret the coeficients: Over the period studied, on average black people experienced 4.00 (exp(1.38849)) times more uses of force per person than white people in any given census tract.

Specification 2: Account for crime rate and racial make up

$$y = n_j e^{\beta_0 + \alpha_j + \beta_{black} X_{black} + \beta_{pov} X_{pov} + \beta_{pctb} X_{pctb} + \beta_{crime} X_{crime}}$$

Where all variables are the same as previous with the addition of X_{pctb} percent of a tract that is black, X_{pov} the percent of the track below the federal poverty line, and X_{crime} the number of reported part 1 (mandated federal reporting) crimes per 100 people in the census track.

Improvements: With pop by race as offset we are predicting number of use of force incidents per person in a race-tact group. It therefore makes sense not to use the absolute number of part 1 crimes per tract as a predictor but instead use the number of part1 crimes per person per tract as a predictor.

```
m2 <- glmer(count ~ race + (1| tract_short) + pct_pov + pct_black + part1_per_capita, offset=log(pop_ra

summary(m2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula:
## count ~ race + (1 | tract_short) + pct_pov + pct_black + part1_per_capita
##    Data: count_data
##  Offset: log(pop_race)
##
##      AIC      BIC   logLik deviance df.resid
##   5548.5   5571.0  -2768.3   5536.5      308
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -11.429  -1.303  -0.159   1.898  39.402
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  tract_short (Intercept) 1.068    1.033
## Number of obs: 314, groups:  tract_short, 157
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -6.299348   0.156771  -40.18  < 2e-16 ***
## raceblack         1.385489   0.019262   71.93  < 2e-16 ***
## pct_pov           0.024466   0.006549    3.74 0.000187 ***
## pct_black        -0.003257   0.003739   -0.87 0.383738
## part1_per_capita  0.088778   0.007496   11.84  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) rcblck pct_pv pct_bl
## raceblack   -0.027
## pct_pov     -0.368 -0.001
## pct_black   -0.243 -0.049 -0.581
## prt1_pr_cpt -0.133 -0.003 -0.468  0.095
```

Interpret coefficients: The effect of being black does not seem to have changed despite the addition of the control variables. On average black people experienced 4.00 (exp(1.385489)) times more uses of force per person than white people in any given census tract on top of the effects from the racial compositon, the number of serious crimes per person reported, and the poverty rate of the census tract.

On average an increase of one part 1 crime per 100 people is associated with a 9.3% (exp(0.089058)) increase in use of force incidents holding all other variables constant (including race) within any given census tract.

Noticeably the tract percent black predictor indicates that a 1% increase share of black people in a census tract was associated with a 0.3% (1 - exp(-0.003193)) decline in use of force incidents. However that estimate was not statistically significant. This goes against both intuition that black people might be either more policed in a predominantly black neighborhood and that they would be more singled out in majority white communities. This does go along with the finding in "An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias"" by Andrew GELMAN, Jeffrey FAGAN, and Alex KISS that fewer stops per violent crime or drug arrest were made in majority black neighborhoods, though the comparison is not really fair because that study used arrests not population as a baseline for stop counts.

Also notably, although the effect of being black did not change between group variance went down from 1.809 in the first model to 1.039 meaning the estimate of the idosyncratic difference in force incidents between census tracts went down, which makes sense because we are accounting some of that variation with the control variables.

Specification 3: Interaction neighborhood wealth and race

$$y = n_j e^{\beta_0 + \alpha_j + \beta_{black} X_{black} + \beta_{pov} X_{pov} + \beta_{pctb} X_{pctb} + \beta_{crime} X_{crime} + \beta_1 X_{pov} X_{black}}$$

Where the newly added coefficent β_1 measure the interaction between race and percent poverty i.e. the differential effect of increasing neighborhood wealth on force against black versus white people.

```
m3 <- glmer(count ~ race + (1| tract_short) + pct_pov + pct_black + part1_per_capita + pct_pov:race, of
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.00901793 (tol =
## 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
summary(m3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula:
## count ~ race + (1 | tract_short) + pct_pov + pct_black + part1_per_capita +
##     pct_pov:race
##    Data: count_data
##  Offset: log(pop_race)
##
##      AIC      BIC   logLik deviance df.resid
##   4937.7   4964.0  -2461.9   4923.7      307
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -9.9175 -1.2360 -0.1406  1.4216 25.0722
##
```

```
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  tract_short (Intercept) 1.063    1.031
## Number of obs: 314, groups:  tract_short, 157
##
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.7298129  0.1577365  -42.66  < 2e-16 ***
## raceblack          2.3114623  0.0417605   55.35  < 2e-16 ***
## pct_pov            0.0409925  0.0065672    6.24 4.32e-10 ***
## pct_black         -0.0054983  0.0037316   -1.47    0.141
## part1_per_capita   0.0874708  0.0074784   11.70  < 2e-16 ***
## raceblack:pct_pov -0.0250624  0.0009882  -25.36  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) rcblck pct_pv pct_bl prt1__
## raceblack   -0.126
## pct_pov     -0.376  0.094
## pct_black   -0.238 -0.041 -0.581
## prt1_pr_cpt -0.131 -0.008 -0.467  0.095
## rcblck:pct_  0.119 -0.894 -0.101  0.022  0.008
## convergence code: 0
## Model failed to converge with max|grad| = 0.00901793 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

Note: there are some convergence issues with the random effects constants. Scaling the data could solve this.

Interpret coefficients: The size of the effect of being black increased significantly. In this model, on average black people experience 10.09 (exp(2.3114623)) times (up from 4.00 times) more uses of force per person than white people in any given census tract on top of effects from the racial compositon, the number of serious crimes per person reported, and the poverty rate of the census tract.

The reason for this change is that in the previous model, we were reporting an average additional force incidents against black people after assuming the increase in uses of force in wealthier areas was evenly distributed between the races. This model measures average effect of being black after acknowledging black and white people will be differently affected in wealthier neighborhoods.

Additionally,the interaction term tells us that the increase in the number of incidents of force grows 2.50% faster as census tracts get richer for black people than for white people . Particularly, we see the model predicts an *decrease* of -2.733108 incidents per 1000 people for a white person moving from a tract at the 25th to 75th percentile of wealth, but an *increase* 9.30473 incidents per 1000 people for a black person moving between the same two tracts.

```
q <- quantile(count_data$pct_pov, c(.25,.75))
mask <- (count_data$pct_pov %in% q)
#count_data[mask,]
preds <- predict(m3, count_data[mask,], type="response") *1000 #* count_data[mask,"pop_race"]
a <- cbind(preds, count_data[mask,c("pct_pov", "part1_per_capita", "race")])
diff(a[2:1, "preds"]); diff(a[4:3, "preds"])
```

```
## [1] -2.733108
```

```
## [1] 9.30473
```

```
a
```

```
##         preds pct_pov part1_per_capita  race
## 68   5.206610    11.2         9.474969 white
## 99   7.939718    38.7        11.682243 white
## 246 39.673876    11.2         9.474969 black
## 277 30.369146    38.7        11.682243 black
```

Not only does this model reveal new insights but it is fits the data much better than the previous models. While the second model decreased model deviance by 162.21 over the first model, the this model decreases model deviance by 775 over the first model. A low deviance is an indication that the model fits the data well.

```
#1- pchisq(deviance(m1) - deviance(m3), attr(logLik(m3), "df") - attr(logLik(m1), "df"))
anova(m1,m2,m3)
```

```
## Data: count_data
## Models:
## m1: count ~ (1 | tract_short) + race
## m2: count ~ race + (1 | tract_short) + pct_pov + pct_black + part1_per_capita
## m3: count ~ race + (1 | tract_short) + pct_pov + pct_black + part1_per_capita +
## m3:     pct_pov:race
##    Df    AIC  BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m1  3 5704.7 5716 -2849.4   5698.7
## m2  6 5548.5 5571 -2768.3   5536.5 162.21      3  < 2.2e-16 ***
## m3  7 4937.7 4964 -2461.9   4923.7 612.81      1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Specification 4: adding murder

I add an additional variable of the number of murders per 1000 people in census tract (across races) and an interaction of that term with being black.

```
m6 <- glmer(count ~ race + (1| tract_short) + pct_pov + pct_black + part1_per_capita + murder_per_capita
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.00461577 (tol =
## 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
summary(m6)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula:
## count ~ race + (1 | tract_short) + pct_pov + pct_black + part1_per_capita +
##     murder_per_capita * race + pct_pov:race
##    Data: count_data
##  Offset: log(pop_race)
##
##      AIC      BIC   logLik deviance df.resid
##   4904.6   4938.4  -2443.3   4886.6      305
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.5857 -1.2046 -0.0778  1.4935 25.0115
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  tract_short (Intercept) 1.043    1.021
## Number of obs: 314, groups:  tract_short, 157
##
## Fixed effects:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -6.7018879  0.1594419  -42.03  < 2e-16 ***
## raceblack                    2.3723281  0.0427872   55.44  < 2e-16 ***
## pct_pov                      0.0403767  0.0065115    6.20 5.62e-10 ***
## pct_black                   -0.0082896  0.0039589   -2.09   0.0363 *
## part1_per_capita             0.0836185  0.0076957   10.87  < 2e-16 ***
## murder_per_capita            0.2326976  0.0941422    2.47   0.0134 *
## raceblack:murder_per_capita -0.0960102  0.0157667   -6.09 1.13e-09 ***
## raceblack:pct_pov           -0.0243117  0.0009931  -24.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) rcblck pct_pv pct_bl prt1__ mrdr__ rcb:__
## raceblack   -0.125
## pct_pov     -0.375  0.090
## pct_black   -0.286 -0.049 -0.530
## prt1_pr_cpt -0.175 -0.014 -0.440  0.181
## mrdr_pr_cpt  0.186  0.051 -0.034 -0.356 -0.268
## rcblck:mr__  0.023 -0.226  0.009  0.025  0.012 -0.152
## rcblck:pct_  0.112 -0.830 -0.101  0.023  0.010  0.006 -0.146
## convergence code: 0
## Model failed to converge with max|grad| = 0.00461577 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

We see that the effect of being black is slightly raised to 10.72 (exp(2.3723281)). Although there are relatively large errors.

We see that model deviance is slightly reduced suggesting a modest increase in fit to the data. The improvement is likely not great because I am adding information that is partly embedding in the part 1 crime variable.

What I would ultimately like to do is change the offset to the part 1 crimes per 100 people variable. This would allow the model to estimate the number of force incidents per reported crime which would be a better baseline understand discrimination for comparable crimes, although obviously the reported crime rate is not the real crime rate and hieghtened police presence can increase the number of crimes found even though the underlying rates are not porportionaly elevated.

Additionally the type of crime or the race of the subject would be good variables to think about in the future. A variable not present but which might be important is the number of years a given police officer has on the job.

Visualizations

Prep data

```
## clean/prep data for visulization

data <- viz_data

# look at the values for each variable
as.data.frame(table(data$race))
```

```
##              Var1  Freq
## 1           ASIAN    21
## 2           BLACK 15095
## 3        HISPANIC    95
## 4         MISSING     1
## 5    MULTI_RACIAL    12
## 6      OTHER_RACE    59
## 7 VALUE_NOT_KNOWN   368
## 8           WHITE  5350
```

```
as.data.frame(table(data$force))
```

```
##                                Var1 Freq
## 1               ACCIDENTAL DISCHARGE   11
## 2                  CHEMICAL IRRITANT 4703
## 3                 INJURY TO PRISONER 6029
## 4                            MISSING    1
## 5      NONCOMPLIANT SUSPECT/ARRESTEE 2466
## 6 TASER-BEANBAG-PEPPERBALL-40MM FOAM 5779
## 7         USE OF FORCE INVESTIGATION 1802
## 8      WEAPON DISCHARGE AT AN ANIMAL  210
```

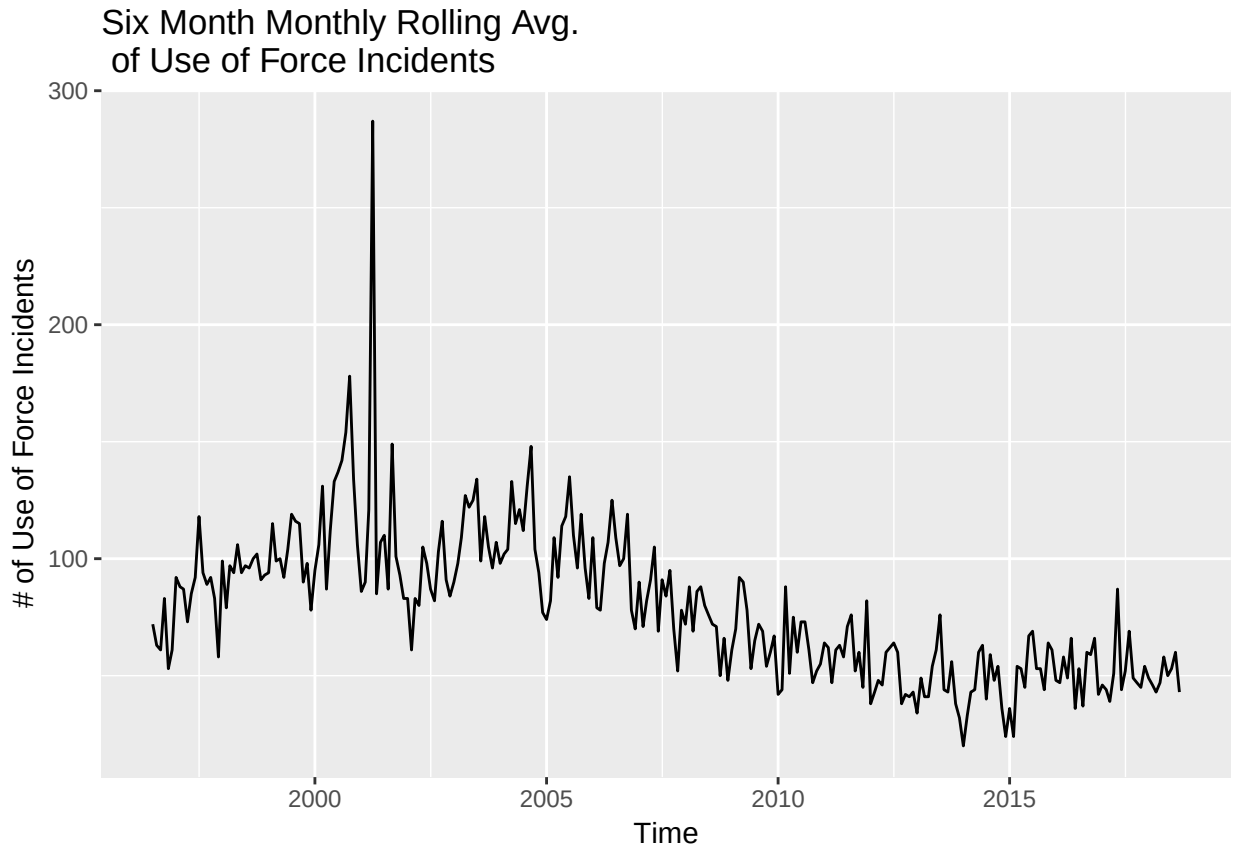
```
# remove missing values
mask <- (data$race!="MISSING") & (data$force!="MISSING") & (!is.na(data$race))
data <- data[mask, ]

# get month
data$date <- as.Date(data$date, format = "%m/%d/%Y")
data$date <- format(data$date, "%m-%Y")
```

Visualization 1

- A line chart showing a rolling average of the number of use of force incidents by month. The window for the rolling average should be 6 months (so the observation for June should be the average of the counts from January-June).

```
# create data
by_month <- data %>% group_by(date) %>% summarise(count = n())
by_month$date <- as.Date(as.yearmon(by_month$date, "%m-%Y"))
by_month <- arrange(by_month, date)
# create rolling avg
by_month$MA <- rollapply(by_month$count, FUN=mean , width=6, align="right", fill=NA)
```

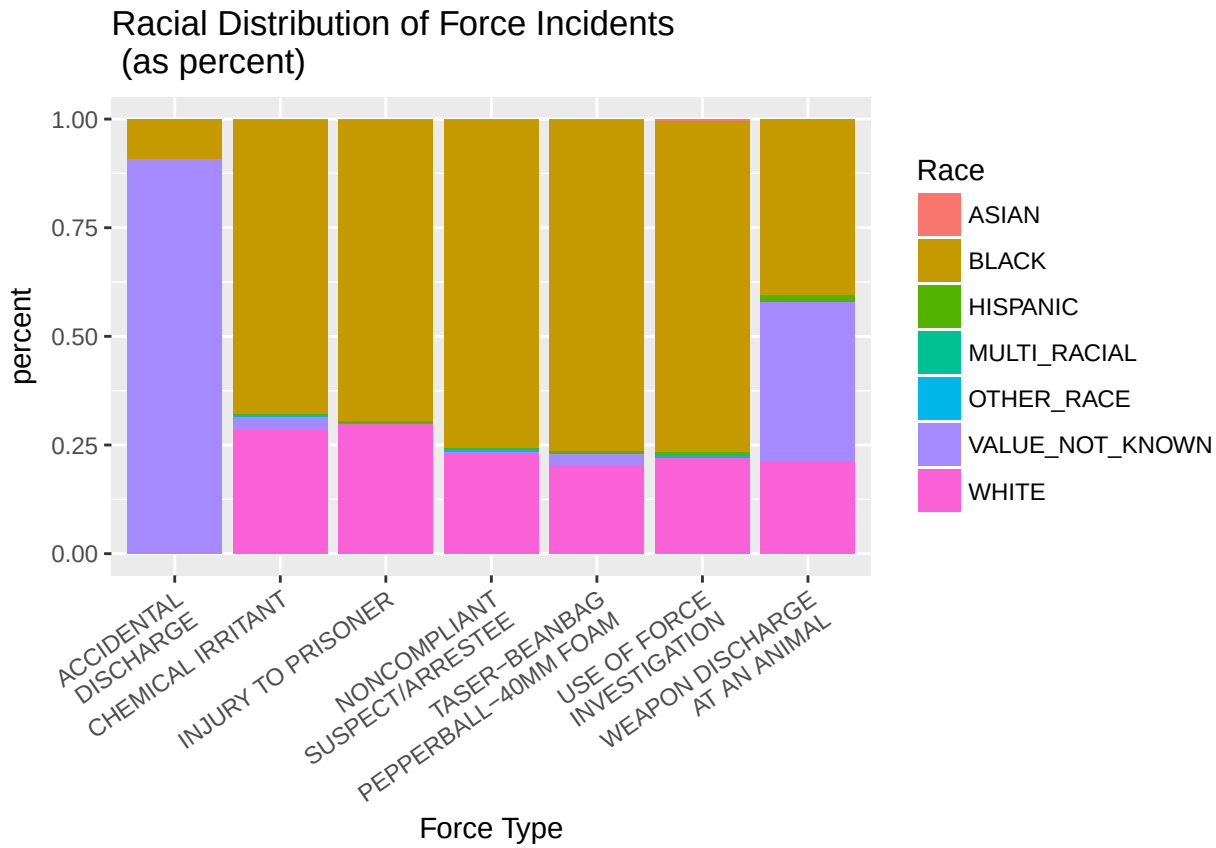
```
ggplot(by_month) +
  geom_line(aes(x=date, y = count)) +
  labs(title="Six Month Monthly Rolling Avg.\n of Use of Force Incidents",
       x="Time", y="# of Use of Force Incidents")
```



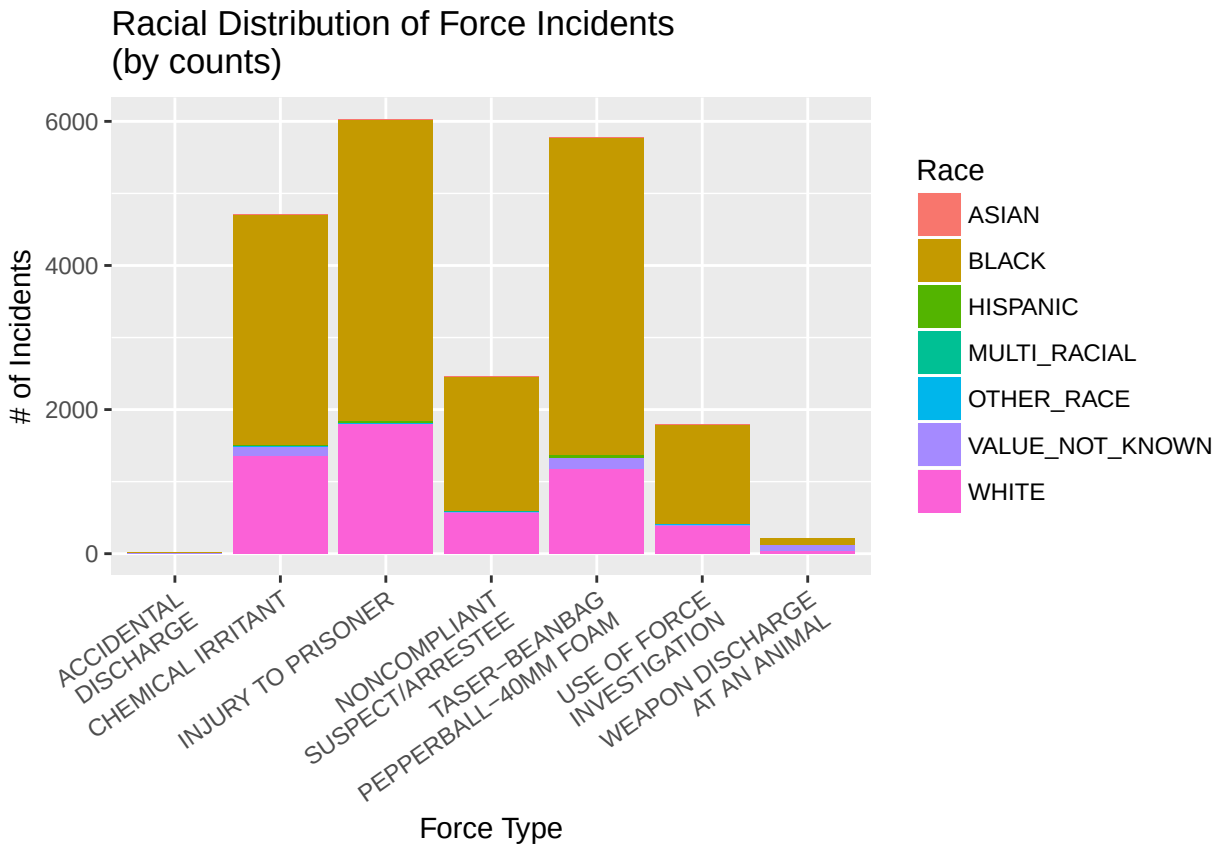
Visualization 2

- A stacked bar chart showing the percentage of use of force incidents by type of force used, broken down by race (one bar for each race, each level in the bar should indicate the type of force used, the thickness of each layer should correspond to the percentage of incidents involving that race and force type; each bar should sum to 100).

```
ticks <- c("ACCIDENTAL\n DISCHARGE",
 "CHEMICAL IRRITANT",
"INJURY TO PRISONER",
 "NONCOMPLIANT\n SUSPECT/ARRESTEE",
"TASER-BEANBAG\nPEPPERBALL-40MM FOAM",
 "USE OF FORCE\n INVESTIGATION",
 "WEAPON DISCHARGE\n AT AN ANIMAL" )

ggplot(data) +
  geom_bar(aes(x=factor(force), fill=factor(race)), position="fill") +
  theme(axis.text.x = element_text(angle= 35, vjust = 1, hjust=1)) +
  scale_x_discrete(labels= ticks) +
  labs(title="Racial Distribution of Force Incidents\n (as percent)", y="percent", x="Force Type", fill=
```




```
ggplot(data) +
  geom_bar(aes(x=factor(force), fill=factor(race))) +
  theme(axis.text.x = element_text(angle= 35, vjust = 1, hjust=1)) +
  scale_x_discrete(labels= ticks) +
  labs(title="Racial Distribution of Force Incidents\n(by counts)", y="# of Incidents", x="Force Type",
```



Visualization 3

- A bar chart showing the number of use-of-force incidents per 1000 residents broken down by race (one bar for each race).

I dropped race categories "VALUE_NOT_KNOWN" and "MULTI_RACIAL" for this visualization. It would be impossible to get the underlying number of people in the area under study who would be categorized "VALUE_NOT_KNOWN". It was difficult to determine in the given time whether "MULTI_RACIAL" overlapped with other racial categories so that was also not included.

```
pop <- viz_pop

mask <- !(data$race %in% c("VALUE_NOT_KNOWN", "MULTI_RACIAL"))
data <- data[mask,]
tracts <- intersect(data$tract_short, pop$tract_short)
a <- pop[pop$tract_short %in% tracts,-c(1:2,8)]
data <- data[data$tract_short %in% tracts,]
# total residents by race
total_pop <- data.frame(totals = colSums(a), race=names(colSums(a))) %>%
  arrange(race)

# get use of force counts by race
race_count <- group_by(data, race) %>% summarise(count = n()) %>% arrange(race)
race_count$race <- tolower(race_count$race)
# get per 1000 in each race
race_count$per_1000 <-  (race_count$count/total_pop$totals) * 1000
```

```
ggplot(race_count) +
  geom_col(aes(x=race, y=per_1000, fill=race)) +
  labs(title="Force Incidents per 1000 People\n by Race", y="# of Incidents", x="Race", fill="Race")
```