Likelihood of Violence at Protests

Peter L'Oiseau

5/12/2021

Motivation

The Mass Mobilization Project has put together a data set spanning the last three decades documenting protests from around the world. More can be read about the project and its funding (1), but the goal of the project is to understand more about citizens' movements against their governments across the world and how those governments respond. Collecting more information about why citizens demonstrate against their governments, how they demonstrate and how governments respond will allow us to systematically assess the political priorities of citizenries around world. And ideally with these assessments governments can formulate advice on how to mitigate tension and violence and create a path to a more harmonious relationship between people and government.

The data has been made open source for researchers to provide more insight into the power that the data holds. It contains information about the location, time, length, purpose, size and certain outcomes of the protest. In this study, I will examine the outcome of violence occurring at a protest. Violence in this study includes beatings, killings and shootings but does not include crowd dispersion tactics or arrests being made. There are distinctions made about whether violence comes from the protesters or the state in response to the protest which is relevant information included in the data but is considered out of scope for this study. In order to predict the likelihood of violence occurring at a protest, we consider where and when the protest took place, what the protesters were demanding, how long the protest was, how many people attended, and how many protests occurred in that country that year. The models used then allow us to understand what countries and regions are more likely to see their protests lead to violence and what other factors contribute to a socially conscious action resulting in violence.

```
#this chunk installs necessary packages for the script and opens the libraries
if (!require("pacman")) install.packages("pacman")
pacman::p_load('RCurl','dplyr','tidyr','kableExtra','purrr','broom','ggplot2','GGally','rsample','lme4','lme
rTest','caret','Metrics','pROC')

library(RCurl)
library(dplyr)
library(tidyr)
library(kableExtra)
library(broom)
library(purrr)
library(ggplot2)
library(GGally)
library(rsample)
library(lme4)
library(lmerTest)
library(caret)
library(Metrics)
library(pROC)
```

Data

The data provided by the Mass Mobilization Project has information from 15,239 distinct protests from 1990 to 2020 in

166 countries. Absent from this set of countries most notably is the United States of America but also includes Israel and some smaller countries. The data has information on the start and end dates of the protest, the number of protests in that country in that year, the identity of the protesters, how many protesters there were, the location, whether the protesters were violent, what demands they made and how the state responded. There are also two additional free text columns which contain sources and notes about the protest. With the protester violence column and the state response columns, I can derive the response variable of interest, protest violence. I can also then derive the length of the protest in days as a potential explanatory variable. Many of the values for number of protesters are variable estimates or ranges and have been simplified in this analysis to the ordinal categories: 50-99, 100-999, 1,000 - 1,999, 2,000 - 4,999, 5,000 - 10,000, and greater than 10,000. While, the notes and sources contain interesting and potentially useful data in free text form, they are considered out of scope. The specific protester identity category and location categories are very expansive but are not used in lieu of country and protester demand information. Note also, there are 4 protester demand columns and 7 state response columns. Not all of these columns are full for all protests since there are a variable number of demands from a protest and state responses to that protest.

```
url <-
  url(
    "https://raw.githubusercontent.com/datacamp/careerhub-data/master/Mass%20Protest%20Data/protest_data.cs
v"
  )
protests <- read.csv(url)

#show data set structure
kable(head(
  protests %>% select(
    country,
    year,
    region,
    protestnumber,
    location,
    protesterviolence,
    participants,
    protesteridentity,
    protesterdemand1,
    stateresponse1
  )
), caption = "Select Variables from the Data Set") %>% kable_styling('striped')
```

Select Variables from the Data Set

country	year	region	protestnumber	location	protesterviolence	participants	protesteridentity	protesterdemand1
Canada	1990	North America	1	national	0	1000s	unspecified	political behavior, process
Canada	1990	North America	2	Montreal, Quebec	0	1000	unspecified	political behavior, process
Canada	1990	North America	3	Montreal, Quebec	0	500	separatist parti quebecois	political behavior, process
Canada	1990	North America	4	Montreal, Quebec	1	100s	mohawk indians	land farm issue
Canada	1990	North America	5	Montreal, Quebec	1	950	local residents	political behavior, process
Canada	1990	North America	6	Kahnawake Reservation near Montreal, Quebec	0	200	mohawk indians	police brutality

Analysis

The analysis plan is to build a binary classification model to predict whether a protest will become violent. I will consider models which have interpretable parameters in order to learn more about the underlying factors which contribute to the likelihood of protests being violent. To do this, I will transform the data into a suitable format for building these models and engineer new features. I will conduct exploratory data analysis on this data to find relationships between the variables and ensure data quality. Next, I will split the data into test and train sets and a further 5-folds of cross-validation in order to compare the efficacy of different models. I will then compare the models across a variety of performance evaluation metrics on the average of the validation sets and determine the best model. I will then train the best model and test it to view it's final efficacy and ultimately analyze the model parameters to come to conclusions and recommendations about what leads to violence at a protest and what can be done.

Performance Metrics

In order to determine which model is the best for understanding the data, I need to first layout an objective criteria which the models will be measured by. Since we have a binary classification problem there are many performance metrics already derived in the literature which we can rely on to grade our models.

Area Under Curve (AUC) of the Receiver Operatering Characteristic (ROC):

This will be our primary metric for determining the efficacy of a model. The ROC is the curve which is defined by varying the cut off point for where one classifies all below predictions as 0 and all above predictions as 1. On the x-axis of the curve is the false positive rate or the percentage of predictions which were labeled false given the threshold when they were actually true and on the y-axis is the recall or the number of predictions labeled true when the actually were divided by the total number of actually true events. After finding this curve for a range of cut off points from 0 to 1, one then calculates the area under the curve. This AUC can range from 0 to 1 but an entirely random predictor would score on average .5 in this metric where 1 is the best score.

Accuracy

Accuracy is simply the number of correct predictions a model made divided by the number of predictions made. This metric can be skewed when there is a large class imbalance between the two outcomes in the data set. However, this data set sees 30.6% of the protests scored as violent which is not an unacceptable level imbalance for relying on accuracy as a metric.

F1

F1 score is metric which balances the precision and recall of a model. Recall is the number of predictions labeled true by the model when the actually were divided by the total number of actually true events, while precision is the number of correctly predicted true events divided by the total number of predicted true events. F1 then is calculated by 2 * (recall * precision)/(recall + precision) which acts as a harmonic mean of recall and precision, two valuable metrics for understanding model performance.

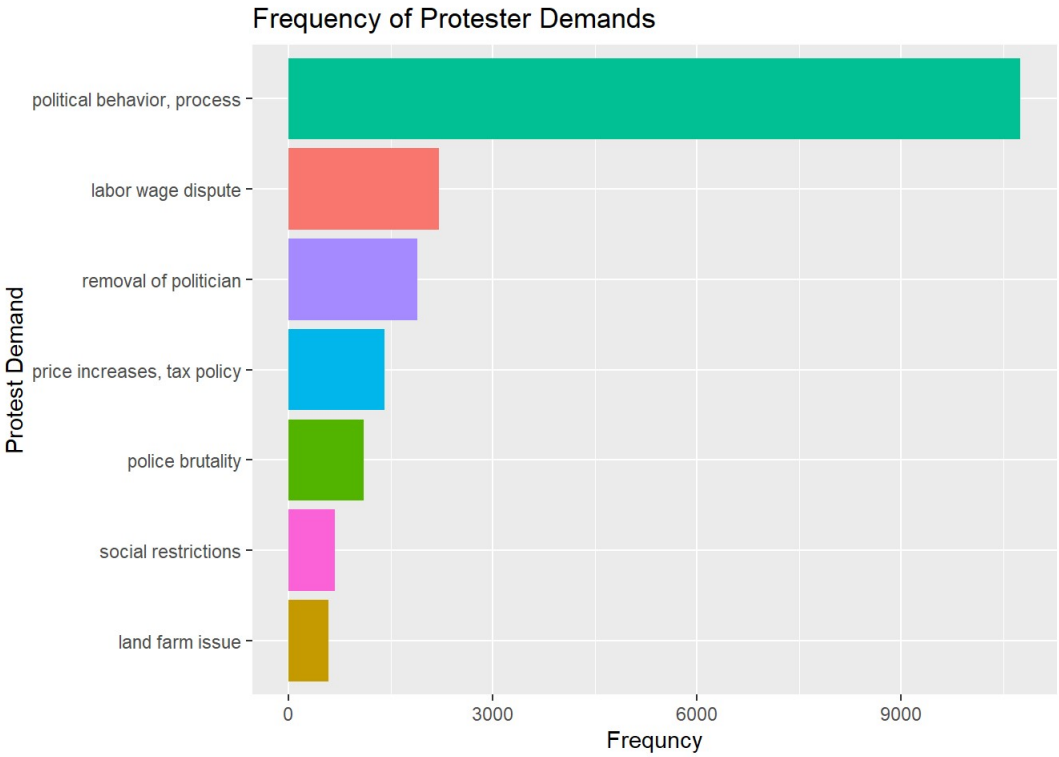
Exploratory Data Analysis

This section will explore the data, the relationships and distributions that lie within and investigate any data errors. You can read more about the data set and its variables here (https://dataverse.harvard.edu /file.xhtml?persistentId=doi:10.7910/DVN/HTTWYL/TJJZNG&version=5.0).

First we explore the 4 protest demand columns to see what the distribution of demands is from the protesters.

```
#get protest length
protests <- protests %>% mutate(
  start_date = as.Date(paste(startyear, startmonth, startday, sep = '/')),
  end_date = as.Date(paste(endyear, endmonth, endday, sep = '/')),
  length_days = as.numeric(end_date - start_date + 1)
)
#note a 1993 protest in Gabon has no demands listed. based on the additional information supplied, I manuall
y categorized it as a political behavior, process
protests[protests$id == 4811993003, 'protesterdemand1'] <- 'political behavior, process'
# 29 protests have no state response listed which will be filled with ignore and 2 are listed only in the se
cond column
protests <- protests %>% mutate(stateresponse1 = ifelse((protest == 1 &
                                                           stateresponse1 == '' &
                                                           stateresponse2 != ''),
                                                        stateresponse2,
                                                        ifelse((protest == 1 &
                                                                  stateresponse1 == '' &
                                                                  stateresponse2 == ''),
                                                               'ignore',
                                                               stateresponse1
                                                        )
))

#put the data into a nested structure to understand protest demands & state responses
protests_nest <-
  protests %>% gather(demand_num,
                      protesterdemand,
                      protesterdemand1:protesterdemand4) %>% filter(
                        protesterdemand !=
                          '' |
                          sources ==
                          '' |
                          (demand_num == 'protesterdemand1' &
                             protesterdemand == '')
                      ) %>% gather(response_num, stateresponse, stateresponse1:stateresponse7) %>% filter(
                        stateresponse !=
                          '' |
                          sources ==
                          '' |
                          (response_num == 'stateresponse1' &
                             stateresponse == '')
                      ) %>% mutate(
                        demand_num = gsub(demand_num, pattern = '[[:alpha:]]', replacement = ''),
                        response_num = gsub(response_num, pattern =
                                              '[[:alpha:]]', replacement = '')
                      ) %>%
  nest(
    -c(
      id,
      year,
      country,
      ccode,
      region,
      protest,
      protesterviolence,
      protestnumber,
      length_days
    )
  )

#distinct protester demands and state responses
protests %>% gather(demand_num,
                    protesterdemand,
                    protesterdemand1:protesterdemand4) %>% filter(
```

```
                      protesterdemand !=
                        '' |
                        sources ==
                        '' |
                        (demand_num == 'protesterdemand1' &
                           protesterdemand == '')
                    ) %>% select(protesterdemand) %>% table() %>% as.data.frame() %>% filter(!(`.` %in%
                                                                                              c('',
'.'))) %>% ggplot(aes(reorder(`.`, Freq), Freq, fill =

`.`)) + geom_col() + coord_flip() + labs(x = 'Protest Demand', y = 'Frequncy', title =

'Frequency of Protester Demands') + theme(legend.position = "none")
```



By far the most frequent demand is with respect to political behavior and process followed by labor wage disputes and removal of a politician. Note, every protest has been categorized into at least one of these seven categories but they are not mutually exclusive. Here is the distribution of the number of demands for a protest.

```
protests %>% gather(demand_num,
                    protesterdemand,
                    protesterdemand1:protesterdemand4) %>% filter(
                      protesterdemand !=
                        '' |
                        sources ==
                        '' |
                        (demand_num == 'protesterdemand1' &
                           protesterdemand == '')
                    ) %>% group_by(id) %>% summarise(n = n()) %>% pull(n) %>%
  table() %>% as.data.frame() %>% kable(caption = 'Frequency of Number of Demands Totals',
                                        col.names = c('Number of Demands', 'Frequency')) %>% kable_styling('
striped')
```

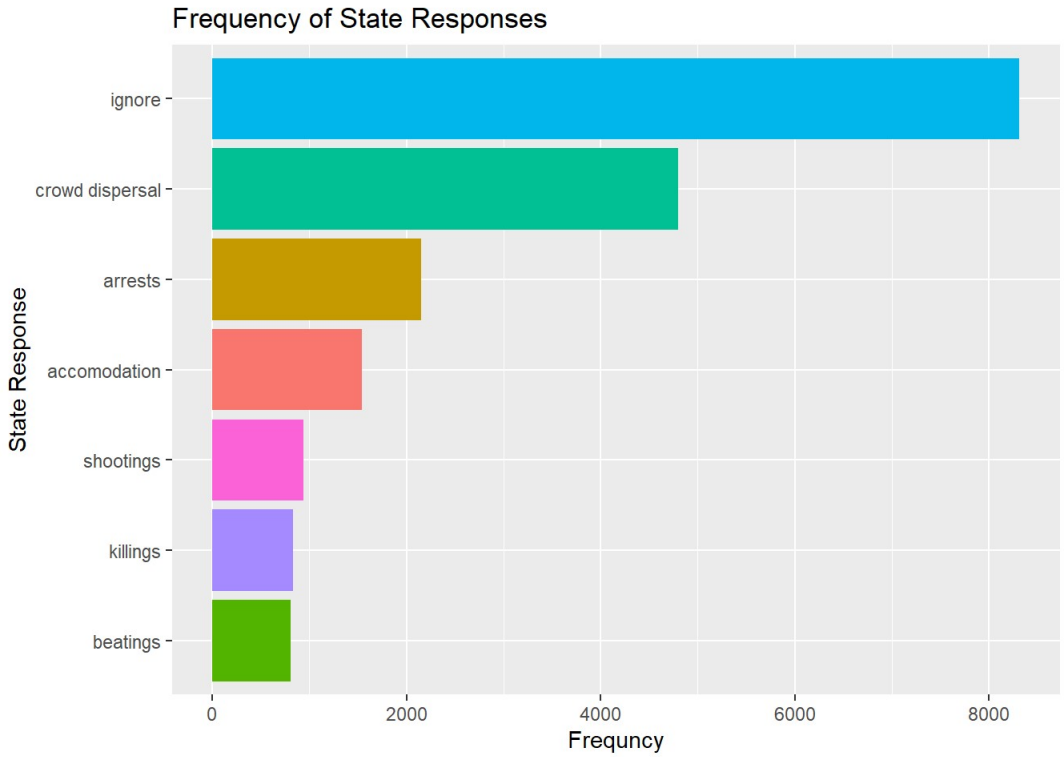
Frequency of Number of Demands Totals

Number of Demands	Frequency
1	11799

Number of Demands	Frequency
2	2777
3	615
4	1954

In light of this variable number of protest demands, I've decided to include the number of demands as a variable in the data and collapse the demands into one column with the seven categories, and have a row for each demand of a protest. Note that this will then result in having multiple predictions for the same protest based on their distinct demands. These predictions will be averaged and used to create a single prediction for each protest. Now we look at the distribution of state responses to these protests.

```
protests %>% gather(response_num, stateresponse, stateresponse1:stateresponse7) %>% filter(
  stateresponse !=
    '' |
    sources ==
    '' |
    (response_num == 'stateresponse1' &
       stateresponse == '')
) %>% select(stateresponse) %>% table() %>% as.data.frame() %>% filter(!(`.` %in%
                                                                          c('', '.'))) %>% ggplot(aes(reord
er(`.`, Freq), Freq, fill =
                                                                                                      `.
`)) + geom_col() + coord_flip() + labs(x = 'State Response', y = 'Frequncy', title =

'Frequency of State Responses') + theme(legend.position = "none")
```



Here again each protest has been categorized as at least one of these responses but can actually have all 7. For defining state violence in response to a protest we take all the protest which have shootings, killings and or stabbings listed as a response. Next we investigate the protester identity column. There are 5,785 distinct identities listed for the 15,239 protest and here are the most frequent.

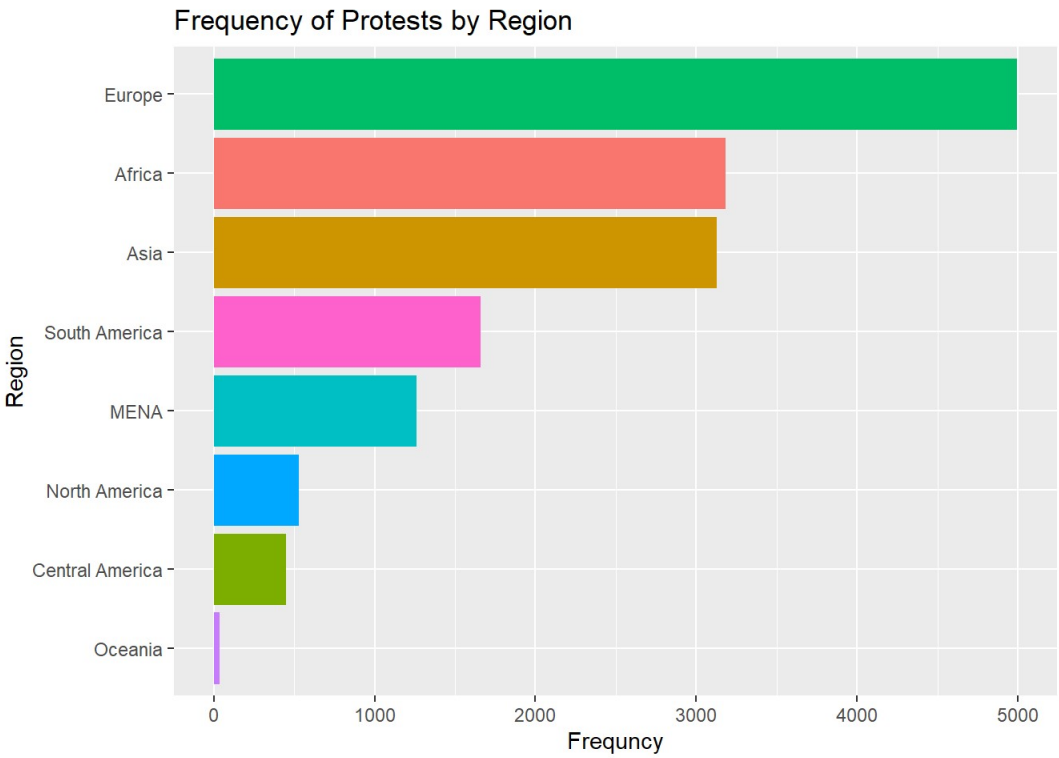
```
#I think a model which attempts to predict which protests will result in protester or state violence is poss
ible.
#Based on the data, I think we can create a mixed effect model which treats country and region and random ef
fects rather than fixed ones.
#I like this approach since we are pooling information within countries and regions which acknowledges the g
eopolitcal reality that people in
#the same geographical area influence social behavior more than people across the world.

#We can use year of protest, length of protest, number of participants, number of demands, what the demands
are,
#number of protests in the country that year and country/region as random effects to understand the likeliho
od of violence

#Note that participant numbers will need to be inspected and manipulated to be a useful factor. Also the pro
tester
#identity is too expansive at this point to be useful but could be categorized perhaps.
as.data.frame(table(protests$protesteridentity)) %>% arrange(desc(Freq)) %>% top_n(10) %>% kable(col.names =
c('Protesters', 'Frequency')) %>%
  kable_styling('striped')
```

Protesters	Frequency
	2461
protesters	1541
students	646
workers	273
unspecified	224
farmers	219
residents	171
opposition supporters	165
university students	144
demonstrators	137

The most frequent categories are blank or protester. It's possible someone with more domain expertise may be able to extract some value from these categorizations but for this study, since the column is broadly distributed and weakly defined, I will not use it to predict violence occurring at a protest.

Countries are also categorized in global regions in this data set which may hold some additional information since countries near each other have social and political effects on one another. Here is the distribution of protests by region. Note MENA describes the Middle East and Northern African countries.

```
#plot the region protest frequencies
ggplot(as.data.frame(table(protests%>%filter(protest==1)%>%pull(region))),aes(reorder(Var1, Freq), Freq, fil
l =
             Var1)) + geom_col() + coord_flip() + labs(x = 'Region', y = 'Frequncy', title =
                                                       'Frequency of Protests by Region') + theme(legend.po
sition = "none")
```



The participants and participants category columns also have a broad range of values but almost all protests have some information regarding crowd size. So, when a range is offered, the top of the range is used to categorize it. The variable present in the data represents an ordinal categorical variables with these ranges: 50-99, 100-999, 1,000 - 1,999, 2,000 - 4,999, 5,000 - 10,000, greater than 10,000. Ultimately, this is the final data set used to predict whether a protest will see protester or state violence.

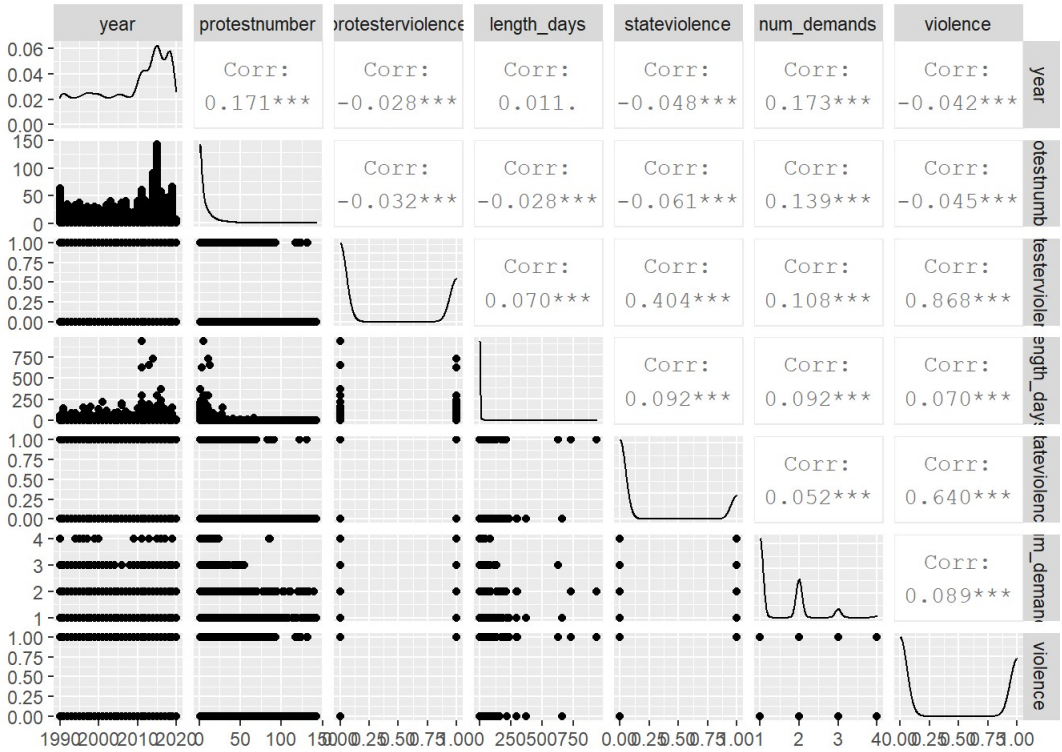
```
#design the dataset to prep for modeling. first eliminate entries for countries with no protest in a given y
ear
protests_nest_filt <- protests_nest %>% filter(protest == 1)

#find out if state response is violent
protests_nest_violence <-
  protests_nest_filt %>% mutate(stateviolence = map_dbl(data,  ~ ifelse(T %in%
                                                                          (.x$stateresponse %in% c('beatings
', 'killings', 'shootings'))
                                                                        , 1, 0)))

#count the number of demands of the protesters
protests_nest_count <-
  protests_nest_violence %>% mutate(num_demands = map_dbl(data, ~ length(unique(
    .x %>% filter(!(protesterdemand %in% c('', '.'))) %>% pull(protesterdemand)
  ))))

#unnest the data to have observations unique to protest and demand, so protest with multiple demands will ha
ve multiple rows
protests_mod <-
  protests_nest_count %>% mutate(data2 = map(
    data,
    ~ .x %>% filter(!(protesterdemand %in% c('', '.'))) %>% select(participants, participants_category, prot
esterdemand)
  )) %>% select(-data) %>% unnest(data2)

#try to categorize the blank participant categories
protests_mod$participants <-
  gsub(protests_mod$participants,
       pattern = 's|\\+|.*-|,|>|<|\\s',
       replacement = '')

protests_mod$participants <-
  gsub(protests_mod$participants,
       pattern = '.*[[:alpha:]]',
       replacement = '')

fin_protests <-
  protests_mod %>% mutate(
    protesterdemand = ifelse(protesterdemand == '.', '', protesterdemand),
    participants = as.numeric(ifelse(
      participants %in% c('', '.'), 0, participants
    )),
    violence = ifelse(protesterviolence == 1 |
                        stateviolence == 1, 1, 0),
    participants_category = ifelse(
      participants_category != '',
      participants_category,
      ifelse(
        participants != 0 & participants >= 50 &
          participants <= 99,
        '50-99',
        ifelse(
          participants >= 100 & participants <= 999,
          '100-999',
          ifelse(
            participants >= 1000 & participants <= 1999,
            '1000-1999',
            ifelse(
              participants >= 2000 & participants <= 4999,
              '2000-4999',
              ifelse(
                participants >= 5000 & participants <= 10000,
                '5000-10000',
```

```
                ifelse(participants > 10000, '>10000', 'unknown')
              )
            )
          )
        )
      )
    )
  )

kable(tail(
  fin_protests %>% select(
    id,
    country,
    year,
    region,
    protestnumber,
    length_days,
    num_demands,
    participants_category,
    protesterdemand,
    violence
  )
)) %>% kable_styling('striped')
```

id	country	year	region	protestnumber	length_days	num_demands	participants_category	protesterdemand
9102016001	Papua New Guinea	2016	Oceania	1	26	1	1000-1999	removal of politician
9102016001	Papua New Guinea	2016	Oceania	1	26	1	1000-1999	removal of politician
9102017001	Papua New Guinea	2017	Oceania	1	1	2	50-99	political behavior, process
9102017001	Papua New Guinea	2017	Oceania	1	1	2	50-99	land farm issue
9102017002	Papua New Guinea	2017	Oceania	2	1	1	50-99	political behavior, process
9102017003	Papua New Guinea	2017	Oceania	3	1	1	100-999	political behavior, process

Note that all the regions have a large number of observations but that is not always the case for individual countries, so, countries with less than 30 observations are pooled together in order to avoid having different levels of factors in the training and testing sets. With this data set ready for modeling, I examine the distribution of the numeric and binary variables and their relationships.

```
#pairs plot of numeric and binary variables
#note that the number of protests and demands are growing over time
ggpairs(fin_protests%>%select(-c(id,ccode,protest,participants))%>%select_if(is.numeric))
```



The number of recorded protests and the number of their demands is increasing over time. The count variable number of protests follows an exponential distribution and the same can be said for the length of the protests in days. Note that protesters and state violence are highly correlated at .4. Here is the confusion matrix describing the observed frequencies of state and protester violence in relation to one another.

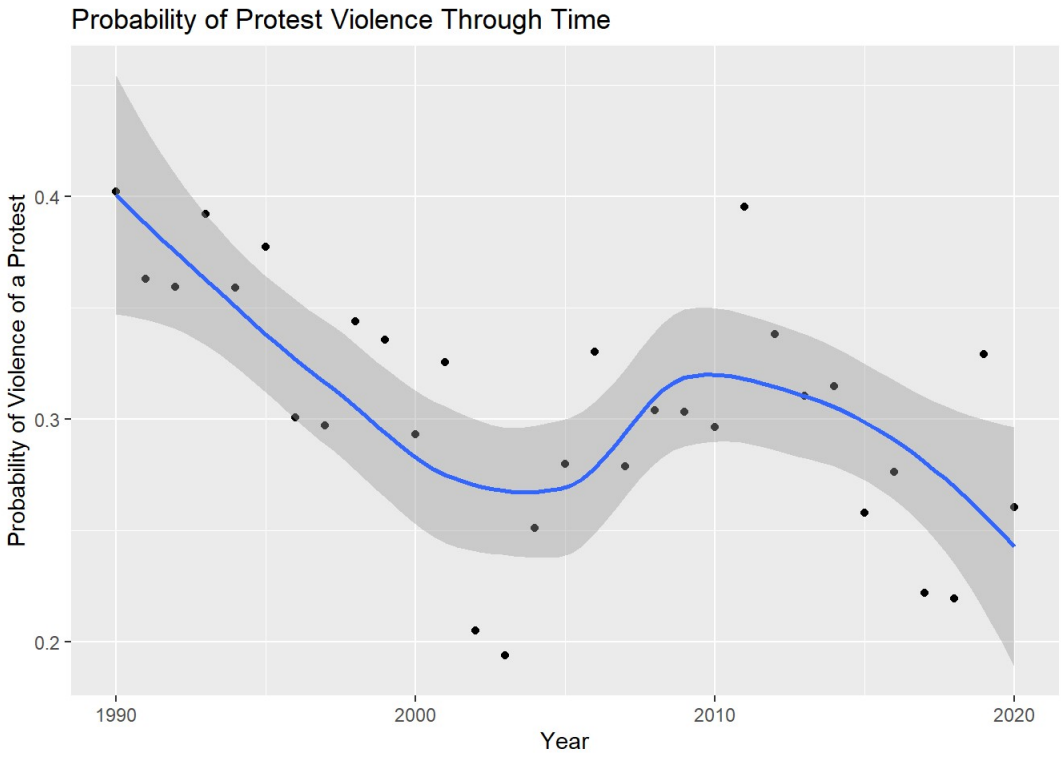
```
#table for protester violence vs state violence
comp_table <- table(
  fin_protests %>% select(id, stateviolence, protesterviolence) %>% distinct() %>%
    select(-id)
)
row.names(comp_table) <-
  c('State Non-Violent', 'State Violent')
kable(
  comp_table,
  caption = 'State Violence compared to Protester Violence',
  col.names = c("Protesters Non-Violent", "Protesters Violent")
) %>% kable_styling('striped')
```

State Violence compared to Protester Violence

	Protesters Non-Violent	Protesters Violent
State Non-Violent	10575	2765
State Violent	629	1270

It appears protester violence without state violence is more common than the inverse however, we will combine the outcomes into violence or no violence as the response variable. Note too how the likelihood of a protest becoming violent has changed over time. In this three decade span there appears to a cubic relationship between time and the likelihood of violence. There may well even be seasonal and periodic time series trends in the data since protest data is a reflection of human behavior which consistently tends to exhibit this sort of statistical behavior. However, the time series approach is not explored in study.

```
#plot how the probability of violence has changed over time
#perhaps a cubic pattern where there are peaks and values over time.
ggplot(
  fin_protests %>% group_by(id, year) %>% summarise(violence = violence[1]) %>%
    group_by(year) %>% summarise(vperc = sum(violence) / n()),
  aes(year, vperc)
) + geom_point() + geom_smooth() + labs(title = 'Probability of Protest Violence Through Time', x = 'Year',
y = 'Probability of Violence of a Protest')
```



With all this in mind, I examine the ability for the data to be modeled with logistic regression. This basic model conducted on the entire data will give us a first pass at understanding the predictive value held in the explanatory variables with respect to the response variable, violence. Here are a selection of the country parameters and the remainder of the explanatory variables and their significance values for this model. Note that many of the variables do indeed appear to have predictive power with respect to the response variable. Protest number is not particularly significant though, with a p-value in excess of .8 so will be discarded from further consideration.

```
#basic logisitic model to predict if violence will occur at a protest
model_df <-
  fin_protests %>% select(-c(
    country,
    protest,
    participants,
    stateviolence,
    protesterviolence
  )) %>% filter(participants_category != 'unknown') %>% group_by(ccode) %>% mutate(ccode = ifelse(n() <
                                                                                                  30, 100
0, ccode))
model_df[, c('id',
             'ccode',
             'violence',
             'region',
             'participants_category',
             'protesterdemand')] <-
  lapply(model_df[, c('id',
                      'ccode',
                      'violence',
                      'region',
                      'participants_category',
                      'protesterdemand')], as.factor)

model_df <-
  model_df %>% mutate_at(vars(year, protestnumber, length_days, num_demands), funs(scale)) %>% mutate_at(var
s(protestnumber, length_days, num_demands),
                                                                                                          ~ replace
(., is.na(.), 0))
model_df$participants_category <-
  factor(
    model_df$participants_category,
    levels = c(
      "50-99",
      "100-999",
      "1000-1999",
      "2000-4999",
      "5000-10000",
      ">10000"
    )
  )

log_model <-
  glm(violence ~ ., data = model_df %>% select(-id) , family = 'binomial')

#evidently there are not enough degrees of freedom for the glm but it shows we do have some predictive power
in our dataset for the outcome variable
kable(tail(tidy(log_model),25),digits=3)%>%kable_styling('striped')
```

term	estimate	std.error	statistic	p.value
ccode850	1.088	0.504	2.159	0.031
ccode910	0.665	0.403	1.650	0.099
ccode1000	1.377	0.367	3.756	0.000
year	-0.149	0.014	-10.304	0.000
regionAsia	-0.568	0.390	-1.458	0.145
regionCentral America	NA	NA	NA	NA

term	estimate	std.error	statistic	p.value
regionEurope	-1.262	0.295	-4.278	0.000
regionMENA	-0.769	0.431	-1.785	0.074
regionNorth America	NA	NA	NA	NA
regionOceania	NA	NA	NA	NA
regionSouth America	NA	NA	NA	NA
protestnumber	-0.003	0.014	-0.179	0.858
length_days	0.216	0.015	13.925	0.000
num_demands	0.164	0.015	11.009	0.000
participants_category100-999	0.083	0.041	2.046	0.041
participants_category1000-1999	0.224	0.056	3.982	0.000
participants_category2000-4999	-0.073	0.051	-1.441	0.150
participants_category5000-10000	0.185	0.059	3.138	0.002
participants_category>10000	0.145	0.050	2.892	0.004
protesterdemandland farm issue	0.431	0.092	4.713	0.000
protesterdemandpolice brutality	0.884	0.071	12.493	0.000
protesterdemandpolitical behavior, process	0.347	0.048	7.218	0.000
protesterdemandprice increases, tax policy	0.626	0.066	9.450	0.000
protesterdemandremoval of politician	0.419	0.063	6.608	0.000
protesterdemandsocial restrictions	-0.483	0.098	-4.922	0.000

Model Development

To begin, let's examine the models that will be considered in this study. We will investigate the utility of logistic regression and compare it to a variety of mixed effect models. First, logistic regression is a classic binary classification model which draws a straight line through the n-dimensional parameter space to try to create the most homogeneous within binary classes with respect to the response variable. This method is not overly complicated but will offer a good baseline of performance and allows us to interpret the parameter estimates for useful recommendations. I will then examine the advantages and drawbacks of a mixed effects model which instead of treating all independent variables as something with a fixed effect on the response variable, we treat certain hierarchical categories as random effects. To treat a variable as a random effect means the acknowledgment that the effect is variable with respect to our response variable. In this case, there are hierarchical categories, country and region which can be treated as random effects instead of fixed ones. This allows the model to pool information across a country or region which reflects the geopolitical reality that protests and other social behavior do not occur in a vacuum but rather spread as people react to others who they see most frequently. I will test three distinct mixed effects models, one which merely uses country and region as random effects, one where these two effects are assumed to be correlated with the year variable and one where they are assumed to be uncorrelated with the year variable. By including the year variable, the model's output for a country's effect of violence at a protest can vary over time

Next, I will separate 75 percent of the observations at random into the training set and the remainder of the data into a test set. I then further split the training set into train and validation sets and repeat that random splitting procedure 4 more times in order to perform 5 fold cross-validation. Each of four models are then trained on all 5 of the training sets separately and then tested on the corresponding validation set. The average across the 5 folds of each model's performance on new data in ROC AUC, accuracy and F1 are then compared below

```
#set seed for reproducibility
set.seed(6629)
#stratify countries to ensure both training and testing set have the same factors
data_split <- initial_split(model_df%>%select(-protestnumber), .75, strata = ccode)

#split into training and testing sets
train <- training(data_split)
test <- testing(data_split)

#perform 5 fold cross-validation
cv_split <- vfold_cv(train, 5, strata = ccode)

#create a tibble to test different models
cv_data <- cv_split %>%
  mutate(train = map(splits,  ~ training(.x)),
         validate = map(splits,  ~ testing(.x)))

#model formulas to try in cross-validation
formula_set <- c(
  violence ~ year + I(year^2) + I(year^3) + length_days + num_demands + participants_category +
    protesterdemand + (year | ccode) + (year | region),
  violence ~ year + I(year^2) + I(year^3) + length_days + num_demands + participants_category +
    protesterdemand + (1 | ccode) + (1 | region),
  violence ~ year + I(year^2) + I(year^3) + length_days + num_demands + participants_category +
    protesterdemand + ccode + region,
  violence ~ year + I(year^2) + I(year^3) + length_days + num_demands + participants_category +
    protesterdemand + (year || ccode) + (year || region)
)

#try glm model, mixed effects model with region and country and then mixed effects model with region and cou
ntry accounting for the covariates of Year
cv_tune <- cv_data %>%
  crossing(formula = formula_set) %>% mutate(
    mod_form = map_chr(formula,  ~ ifelse(
      !(T %in% grepl(.x, pattern = '\\|')), 'logistic', ifelse(T %in% grepl(.x, pattern = '\\(year\\
s\\|\\|'), 'uncorrealated random year', ifelse(T %in% grepl(.x, pattern = '\\(1'),'random no year','correala
ted random year')
    ))),
    mixed = map_lgl(formula,  ~ T %in% grepl(.x, pattern = '\\|')),
    model = pmap(list(
      x = train, y = formula, z = mixed
    ), function(x, y, z)
      if (z) {
        glmer(y, data = x, family = 'binomial')
      }
    else{
      glm(y, data = x, family = 'binomial')
    })
  )

#this tests each fold an models predictions on the validation set
#note that there are multiple entries for the same protest, so i take the mean prediction per id to come to
a conclusion about a protest's prediction
cv_prep_tunerf <- cv_tune %>%
  mutate(
    validate_actual = map(
      validate,
      ~ .x %>% group_by(id) %>% summarise(violence = violence[1] == 1) %>% arrange(id) %>%
        pull(violence)
    ),
    validate_predicted = map2(
      .x = model,
      .y = validate,
      ~ data.frame(pred = predict(.x, .y, type = "response"),
```

```
                  id = .y$id) %>% group_by(id) %>% summarise(pred = mean(pred) >= .5) %>% arrange(id) %>%
        pull(pred)
    ),
    validate_predicted_num = map2(
      .x = model,
      .y = validate,
      ~ data.frame(pred = predict(.x, .y, type = "response"),
                  id = .y$id) %>% group_by(id) %>% summarise(pred = mean(pred)) %>% arrange(id) %>%
        pull(pred)
    )
  )

#Calculate validate recall, precision and F1 for each fold and hyperparameter combination.
#Recall and precision must be balance in order for this model to be useful
#therefore the model with the highest mean F1 value will be chosen as the best model
fcv_eval_tunerf <- cv_prep_tunerf %>%
  mutate(
    validate_recall = map2_dbl(
      .x = validate_predicted,
      .y = validate_actual,
      ~ recall(actual = .y, predicted = .x)
    ),
    validate_precision = map2_dbl(
      .x = validate_actual,
      .y = validate_predicted,
      ~ precision(actual = .x, predicted = .y)
    ),
    validate_roc = map2_dbl(.x = validate_predicted_num, .y = validate_actual, ~
                              roc(.y, .x)$auc),
    validate_accuracy = map2_dbl(.x = validate_predicted, .y = validate_actual, ~
                                   accuracy(.y, .x))
  )

# Calculate the mean F1 for each hyperparameter combination
hyper_par <- fcv_eval_tunerf %>%
  group_by(mod_form) %>%
  summarise(
    mean_recall = mean(validate_recall, na.rm = T),
    mean_precision = mean(validate_precision, na.rm = T),
    mean_roc = mean(validate_roc, na.rm = T),
    mean_accuracy = mean(validate_accuracy, na.rm = T)
  ) %>% mutate(mean_F1 = 2 * (mean_precision * mean_recall) / (mean_precision + mean_recall)) %>% arrange(de
sc(mean_roc))

kable(
  hyper_par %>% select(mod_form, mean_roc, mean_accuracy, mean_F1),
  col.names = c("Model", "ROC AUC", "Accuracy", "F1"),
  digits = 3
) %>% kable_styling('striped')
```

Model	ROC AUC	Accuracy	F1
correalated random year	0.755	0.704	0.597
uncorrealated random year	0.755	0.704	0.597
random no year	0.740	0.691	0.576
logistic	0.739	0.691	0.579

To three decimal places, the correlated year and random effects model and uncorrelated year and random effect model perform similarly in all the performance metrics but very slightly the correlated model performs the best in ROC AUC and accuracy. This demonstrates that treating the country and region of a protest as random effect and acknowledging

that it changes with respect to time is a better way to understand how likely a protest is to become violent. It means that pooling information about protest violence in a country or a region is wise when trying to predict the likelihood of future violence.

Now with the determination of the best model, I fit the model on the entire training set and view the results on the test set.

```
best_model <-   glmer(
  violence ~ year + I(year ^ 2) + I(year ^ 3) + length_days + num_demands + participants_category +
    protesterdemand + (year | ccode) + (year | region),
  family = 'binomial',
  data = train
)

#note the number of demands of the protesters increases the likelihood of violence as well as the length of
the protest
#when used as linear effect protest violence appears to be going down overtime across the world over time bu
t there is
#an individual year effect for each country which is different over time
kable(summary(best_model)$coefficients,digits=3)%>%kable_styling('striped')
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.917	0.224	-4.092	0.000
year	-0.212	0.053	-4.037	0.000
I(year^2)	-0.011	0.019	-0.555	0.579
I(year^3)	-0.002	0.005	-0.501	0.616
length_days	0.222	0.018	12.538	0.000
num_demands	0.140	0.018	7.731	0.000
participants_category100-999	0.069	0.048	1.426	0.154
participants_category1000-1999	0.262	0.067	3.927	0.000
participants_category2000-4999	-0.107	0.061	-1.758	0.079
participants_category5000-10000	0.108	0.072	1.503	0.133
participants_category>10000	0.130	0.060	2.163	0.031
protesterdemandland farm issue	0.545	0.107	5.088	0.000
protesterdemandpolice brutality	0.890	0.084	10.590	0.000
protesterdemandpolitical behavior, process	0.389	0.057	6.856	0.000
protesterdemandprice increases, tax policy	0.607	0.078	7.747	0.000
protesterdemandremoval of politician	0.350	0.076	4.605	0.000
protesterdemandsocial restrictions	-0.517	0.116	-4.464	0.000

Interestingly, the inclusion of the quadratic and cubic terms for the year variable are not considered significant by p-value, but they do improve the ability of the model to make predictions on new data.

```
predictions <-
  as.factor(
    data.frame(
      pred = predict(best_model, test, type = "response"),
      id = test$id
    ) %>% group_by(id) %>% summarise(pred = mean(pred) >= .5) %>% arrange(id) %>%
      pull(pred)
  )

actual <-
  as.factor(
    test %>% group_by(id) %>% summarise(violence = violence[1] == 1) %>% arrange(id) %>%
      pull(violence)
  )

perf_df <- data.frame(
  ROC = as.numeric((
    roc(
      actual,
      data.frame(
        pred = predict(best_model, test, type = "response"),
        id = test$id
      ) %>% group_by(id) %>% summarise(pred = mean(pred)) %>% arrange(id) %>%
        pull(pred)
    )$auc)
  ),
  Accuracy = accuracy(predictions, actual),
  F1 = 2 * (
    caret::recall(predictions, actual) * caret::precision(predictions, actual)
  ) / (
    caret::recall(predictions, actual) + caret::precision(predictions, actual)
  )
)

kable(perf_df, digits = 3) %>% kable_styling('striped')
```

ROC	Accuracy	F1
0.739	0.695	0.759

Note there is some slippage from the validation metrics to the test metrics but regardless, it does appear to be an effective model for understanding the data. We can observe the average effect a region has on the likelihood of violence in the median year of the sample, 2007.

```
region_coefs <-
  as.data.frame(ranef(best_model)$region) %>% cbind(data.frame(region = rownames(ranef(best_model)$region)))
%>%
left_join(protests %>% filter(protest==1) %>% group_by(region) %>% summarise(n=n()))

#note that these score vary over the year in a linear manner
kable(
  region_coefs %>% mutate(Estimate = `(Intercept)` + median(scale(protests$year))*year)%>% select(region, n,
Estimate) %>% na.omit() %>% arrange(desc(Estimate)),
  col.names = c('Region', 'Total Number of Protests','Effect on Liklihood of Violence'), caption = 'Region E
ffect on Protest Violence in Medain Year (2007)'
) %>% kable_styling('striped')
```

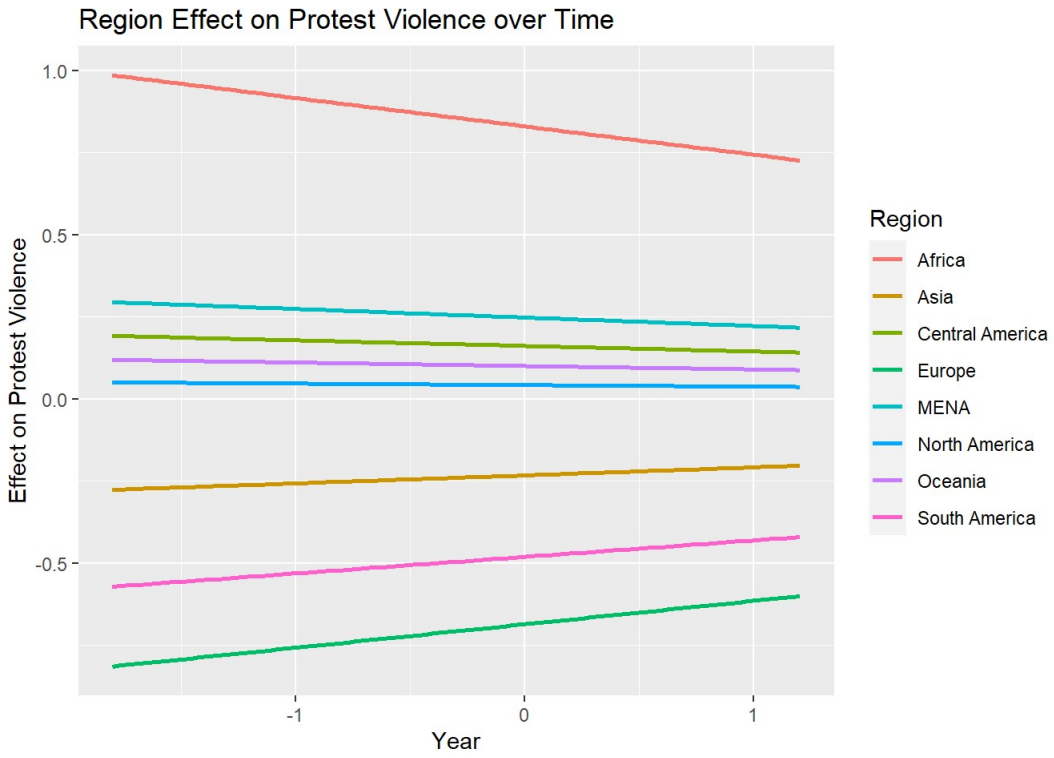
Region Effect on Protest Violence in Medain Year (2007)

Region	Total Number of Protests	Effect on Liklihood of Violence

Region	Total Number of Protests	Effect on Liklihood of Violence
Africa	3184	0.8221945
MENA	1260	0.2457337
Central America	451	0.1600585
Oceania	38	0.0994151
North America	527	0.0418205
Asia	3126	-0.2305097
South America	1659	-0.4764553
Europe	4994	-0.6798384

Evidently, countries in Africa and the Middle East are more likely to see a protest be violent as compared to other regions like Europe and South America. Note again, the exceptions in the data of the United States of America and Israel which may change these estimates if they were included. The model here assumes that region and time are inversely correlated, so a region with a high positive effect on violence at a protest will see that diminish over time and ones with negative effects will increase. This serves as mode of regressing the effect to the global mean and does offer slightly better performance as seen in cross-validation.

```
#countries with more protest violence see it rising over time while others with less see it decreasing
#this is an assumption of the model that year and country effect on violence are perfectly negatively correl
ated
#and performs better than the model which assumes they are uncorrelated
ggplot(
  region_coefs %>% crossing(year2 = min(scale(protests$year)):max(scale(protests$year))) %>%
    mutate(estimate = `(Intercept)` + year * year2) ,
  aes(year2, estimate, col = region)
) + geom_smooth(method='lm',se=F) +labs(title = 'Region Effect on Protest Violence over Time', x = 'Year', y
= 'Effect on Protest Violence', col = 'Region')
```



Lastly, here are the countries with the most negative effect on protest violence (least likely to see violent protests).

```
#also note that some of these countries don't exist at certain points in time (Yugoslavia 1992 disintegratio
n and once more in 2003)
coefs <- as.data.frame(ranef(best_model)$ccode)
protests$ccode <- as.character(protests$ccode)
country_coefs <-
  coefs %>% cbind(data.frame(ccode = rownames(coefs))) %>%
  left_join(protests %>% na.omit() %>% group_by(ccode) %>% mutate(n=n()) %>% select(ccode, country, n) %>% d
istinct())

kable(
  country_coefs %>% mutate(Estimate = `(Intercept)` + median(scale(protests$year)) *
                           year) %>% select(country, n, Estimate) %>% na.omit() %>% arrange(Estimate) %>%
top_n(10, -Estimate),
  col.names = c(
    'Country',
    'Total Number of Protests',
    'Effect on Liklihood of Violence'
  ),
  caption = 'Country Effect on Protest Violence in Medain Year (2007)'
) %>% kable_styling('striped')
```

Country Effect on Protest Violence in Medain Year (2007)

Country	Total Number of Protests	Effect on Liklihood of Violence
Cuba	99	-2.545317
Latvia	85	-1.992193
Uruguay	68	-1.898093
Japan	59	-1.895316
Ireland	431	-1.790304
Serbia	31	-1.640486
Namibia	225	-1.585908
Benin	59	-1.516715
Slovak Republic	44	-1.411573
Mongolia	50	-1.318157

There are countries from all around the world on this list, so it is important to include this alongside the region as a random effect in the model.

Conclusion and Recommendations

Having seen the results of this mixed effects model, one should be confident that when and where a protest occurred can help to predict whether or not a protest is likely to be violent either from the protester side, state side or both. We also see that the longer a protest goes on and the more demands a group is calling for increases the likelihood of violence. The size of the protest is also relevant but the scattered reports for this variable made it difficult to determine the relationship with protest violence, however it appears small crowds below 100 people are significantly less likely to be violent. Protester demands are also important to know when predicting violence as some protests over topics like police brutality are more likely to see violence occur than a protest over social restrictions.

Lastly, let's examine the importance of the random effects of country and region. These are interesting because they suggest that the effect of a region or country are not fixed and can change especially in relation to time. So, there are differences between protest violence in different countries and regions but they do change and ideally this would mean that they can be changed by human intervention. I.e. there's nothing inherent and immutable in the African region land mass that makes it more likely to see protest violence or Cuba less likely. I think an important next step to understanding what governments can do to facilitate a more peaceful culture of protest in their countries is to

investigate what social, political and cultural variables can explain the random variation in the effect a country or region has on protest violence. For example, a researcher could investigate whether Gross Domestic Product (GDP) per capita or government structure of a country is useful in describing the variation in the country effect. If we can understand more about the underlying factors about a country or region which makes it more susceptible to protest violence, then it becomes easier to make effective recommendations on how the violence can be curtailed. Additionally, based on this study, we can say that protesters which are making multiple demands, are willing to protest multiple days and weeks and are assembled in larger numbers are more likely to have violent protests. This knowledge can be used to triage more resources to protests which are at higher risk of becoming violent. Those resources could be police forces, diplomats, social workers or other government officials but more research should be done to understand how a government can effectively reduce the likelihood of violence at ones which are known to be high risk for that outcome.

References

(1)

59 Clark, David Regan, Patrick David H. Clark Political Instability Task Force / CIA

2016-01-13

2016 V5 protests demonstrations event Harvard Dataverse https://doi.org/10.7910/DVN/HTTWYL (https://doi.org/10.7910/DVN/HTTWYL) doi/10.7910/DVN/HTTWYL