Media Bias in Mass Shooting Coverage

Thyne Boonmark, Justin Chen

Abstract Mass shootings and media reporting have been widely discussed topics recently. Media outlets have been accused of using shooting strategies to push some sort of agenda, and this is the topic that we want to explore. When we compared shootings reported by online media outlets to a more comprehensive dataset collected from law enforcement and government records, we found that there were biases in media reporting from political affiliation. Either mass shootings in democratic congressional districts were underreported or those in republican congressional districts were overreported. We also found a bias in media reporting based on time of year. Shootings in the beginning of the year appear to be reported more often than shootings near the end of the year. After observing these possible symptoms of reporting bias, we made a logistic regression model to predict whether or not a mass shooting would be reported by the media. While our model is not perfect, it has an AUC value of 0.7611 which means that the variables we chose are pretty strong predictors.

Introduction Five years ago, on December 14, 2012, Adam Lanza killed 20 school children at the Sandy Hook Elementary School. This horrific mass shooting was covered by the mass media, informing all of American of the shooting. Recently, on October 1, 2017, there was another mass shooting, resulting in 58 deaths and 546 injured. Mass shootings like these have become more common and many are widely covered by the media. However, what about shootings that do not get any media publicity at all? What factors make a shooting more likely to be reported?

The main question we wanted to answer was: "How does the political affiliation of a town affect whether or not a mass shooting there is reported in the media?". The topic of "fake news" and media bias in reporting has been getting alot of discussion, so we wanted to investigate if there exists a tendency for media outlets to report on shootings in a biased way. To perform this analysis we needed two types of datasets. One would have to be a comprehensive list of mass shootings; the other would have to be a list of media reported mass shootings.

Afterwards, we would compare which shootings were reported based on their location's political party affiliation. To determine a town's political party, we would map a location to its congressional district. Some of the code we used to find to find the longitude and latitude values of a city, state.

```
geocodeAdddress <- function(address,which) {
  require(RJSONIO)
  url <- "http://maps.google.com/maps/api/geocode/json?address="
  url <- URLencode(paste(url, address, "&sensor=false", sep = ""))
  x <- fromJSON(url, simplify = FALSE)
  if (x$status == "OK") {
    out <- c(x$results[[1]]$geometry$location$lng,
             x$results[[1]]$geometry$location$lat)
  } else {
    out <- NA
  }
  Sys.sleep(0.2)  # API only allows 5 requests per second
  if (which==0){
    out[1]
  }else{
    out[2]
  }
}
```

Here is some of the code we used to find the party affiliation of some congressional districts.

```r
#Helper method for cleaning
removeEdit <- function(stateText) {
  stateText <- substr(stateText,1, nchar(stateText) - 6)
}

url113 <- "https://en.wikipedia.org/wiki/113th_United_States_Congress"

text113 <- url113 %>% read_html() %>% html_nodes("td") %>% html_text()
text113C <- paste(text113[93], text113[94], sep= "\n")

#Initialize dataframe
dataCongress113 = data.frame("State" = 1:550 , "District" = 1:550,"Party"=1:550, "CongressNum" = 113)

congress113 <- text113C %>%
  strsplit(split = "\n") %>%
  magrittr::extract2(1)

#Cleaning
statesWEdit113<- grep("edit", congress113, value = TRUE)
statesWEditLineNum113<- grep("edit", congress113, value = FALSE)

statesWOEdit113 <- lapply(statesWEdit113, FUN = removeEdit)

#Begin filling data frame with states
for (i in 1:50){
  beg <- statesWEditLineNum113[i] + 1
  end <- statesWEditLineNum113[i+1] - 1
  dataCongress113$State[beg:end] <- statesWOEdit113[[i]]
}

#Congressional districts for each state
for(i in 1:53){
  ndx <- grep(as.character(i), congress113, value=F)
  dataCongress113$District[ndx] <- i
}

#Giving a Party value for the indexes
repIndex <- grep("\\(R\\)", congress113, value = F)
dataCongress113$Party[repIndex] <- "Republican"

demIndex <- grep("\\(D\\)", congress113, value = F)
dataCongress113$Party[demIndex] <- "Democrat"

#Only one district for at large
indexAtLarge <- grep("At-large", congress113, value = F)
dataCongress113$District[indexAtLarge] <- 1

#Special Cases Due to inconsistencies in format
dataCongress113$District[3] <- 1
dataCongress113$District[112] <- 13
dataCongress113$District[154] <- 2
dataCongress113$District[203] <- 5
dataCongress113$District[224] <- 5
```

```
dataCongress113$District[270] <- 8
dataCongress113$District[288] <- 1
dataCongress113$District[346] <- 12
dataCongress113$District[404] <- 1
dataCongress113$District[474] <- 7
```

The second question we wanted to answer was: "Does time of year have any effect the reporting of mass shootings?". We were interested in investigating a possible "media fatigue" phenomenon, where mass shootings would be covered less as people get tired of hearing about it later in the year. We hypothesize that perhaps the media would cover mass shootings cyclically, after people had not heard about one for a while, so that they would generate more readers. To answer this question we needed to compare two different distributions, one for media reported mass shootings and one for all mass shootings. We decided that time would be counted in weeks due to some variances in shooting date across sources.

Here is some of the code we used for creating a density plot

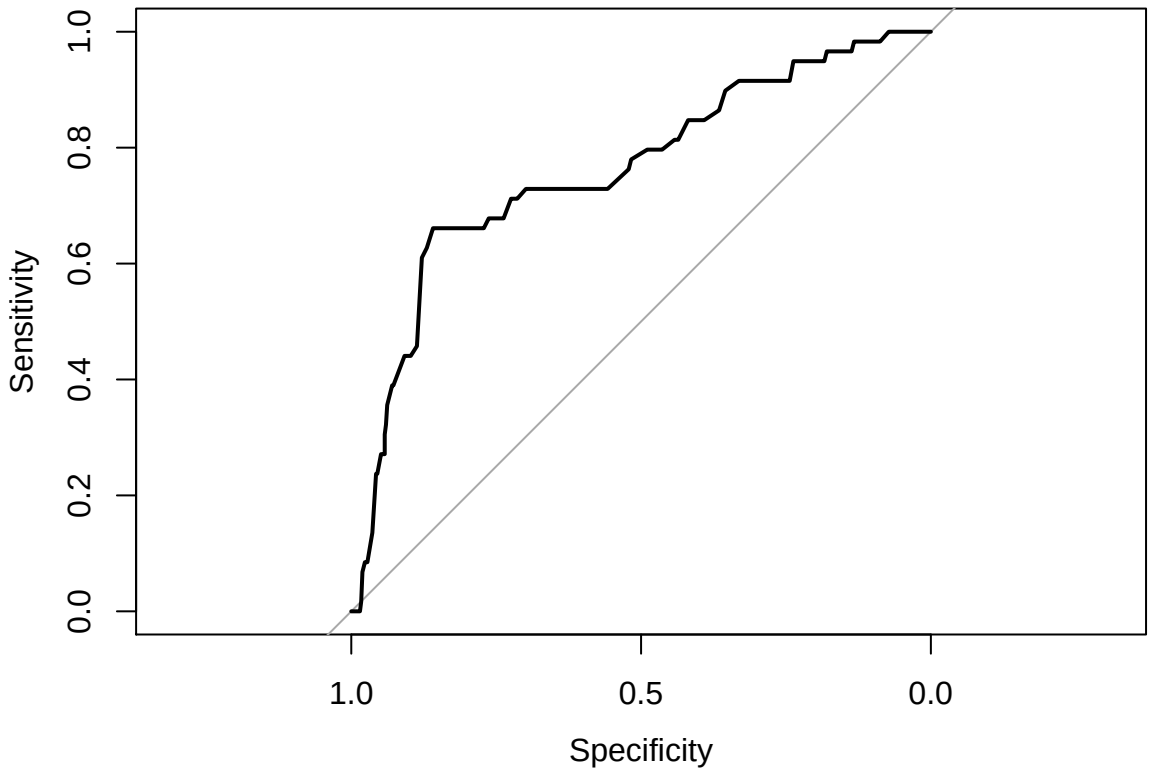
```
shootingFreq2014 <- mass2014 %>%
  group_by(week) %>%
  summarize(numShootings = n())

shootingFreq2015 <- mass2015 %>%
  group_by(week) %>%
  summarize(numShootings = n())

shootingFreq2016 <- mass2016 %>%
  group_by(week) %>%
  summarize(numShootings = n())

shootingFreqAll <- joinedAllFinal %>%
  group_by(week) %>%
  summarize(numShootings = n())

shootingFreqStanford <- joinedStanfordPresentFinal %>%
  group_by(week) %>%
  summarize(numShootings = n())
```

After including and these variables, we wanted to predict whether or not a mass shooting would be reported by the media. To do this, we used a logistic regression model as our predictive model. We split the data into training and test sets, then used political affiliation of the town and week of the year the shooting happened in as variables.

Here is some of the code we used for creating a logistic regression model

```
#creating the log model

#first split into train and test
n <- nrow(joinedAllFinalInclusionTemp)
set.seed(100)
shooting_index <- sample(1:n, round(n*.40))
train <- joinedAllFinalInclusionTemp[shooting_index,]
test <- joinedAllFinalInclusionTemp[-shooting_index,]

#Learn Log model using train set

mod <- glm( IncludedNo ~ as.factor(Party) + week , data=train)
```

```
#make predictions
test$p_hat <- predict(mod, test, type='response')

#compare
roc_obj <- roc(test$IncludedNo, test$p_hat)

plot(roc_obj)
```




```
auc(roc_obj)
```

```
## Area under the curve: 0.7611
```

Data We worked with two datasets covering mass shootings, the Mass Shootings in America dataset collected by Stanford and a Kaggle dataset collected by the Gun Violence Archive. Both of these datasets contain date and location information about their shootings, as well as information on the number of victims killed and the number of victims injured. Date was given in Year/Month/Day format, and location was given both as town and state name (ex: Austin, Texas).

We later added more variables for latitude and longitude, week, congressional district, and congressional district's political affiliation. Latitude and longitude data was collected by using the google maps API to convert our location data into latitude and longitude coordinate pairs. We used R's lubridate package to convert the dates we were given into R date objects and then converted those into weeks. Congressional district information was collected by searching up a given town's district on govtrack.us and Congressional district's political affiliation was collected by scraping data from wikipedia.

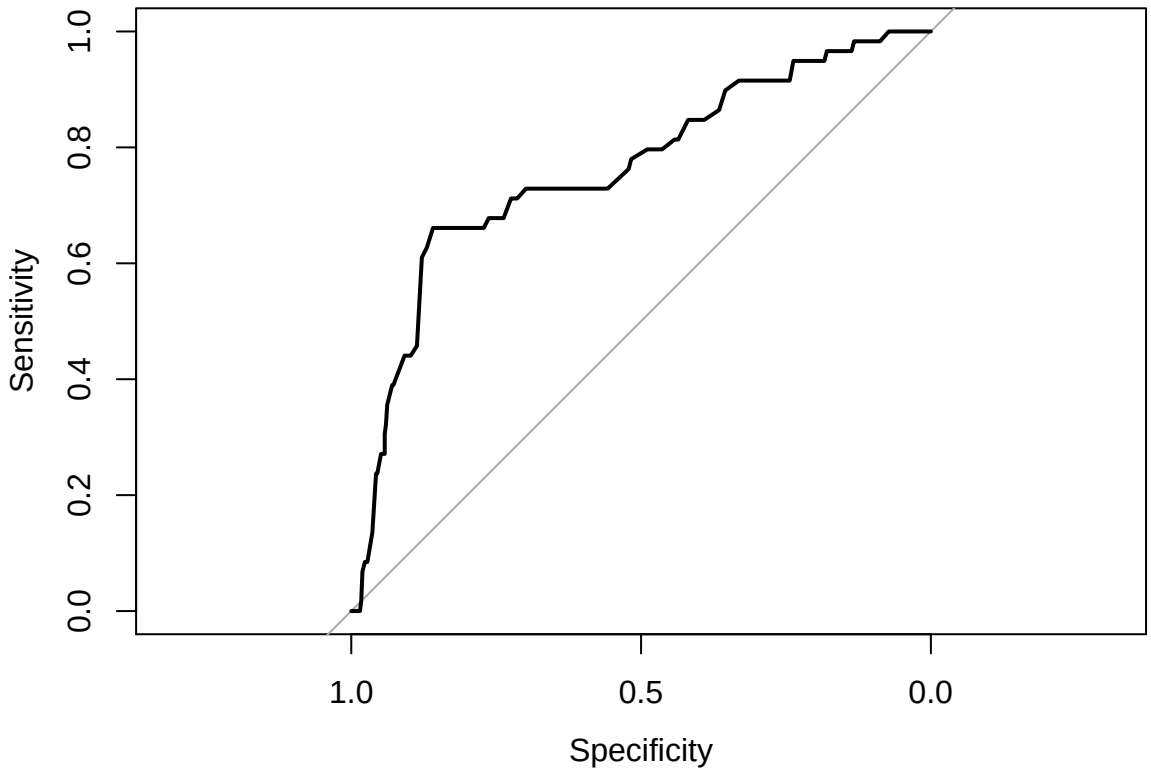
Here is the code we used to join the dataframes together.

```
stanfordCongressDistricts <- read.csv("Congressional Districts Stanford - Sheet1.csv")
allCongressDistricts<- read.csv("Congressional Districts - Sheet1.csv")
totalCongressDistricts <- rbind(stanfordCongressDistricts, allCongressDistricts)
```

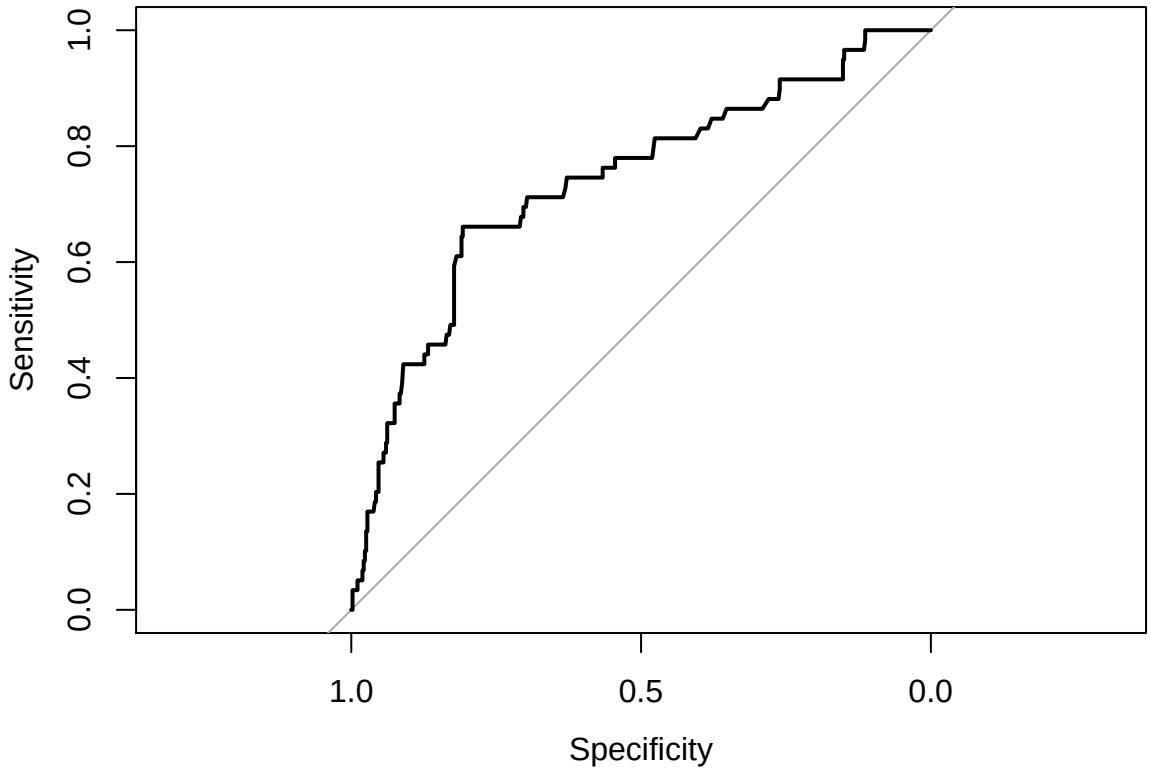
```r
joinedStanford <- inner_join(massStanfordPresent, totalCongressDistricts, by = c("City","State"))

joinedStanfordPresent <- left_join(joinedStanford, dataCongress113, by = c("District", "State", "Congres
joinedStanfordPresentFinal <- left_join(joinedStanfordPresent, dataCongress114, by = c("District", "Sta

for (i in 1:140){
  if (is.na(joinedStanfordPresentFinal$Party.x[i])) {
    joinedStanfordPresentFinal$Party.x[i] <- joinedStanfordPresentFinal$Party.y[i]
  }
}

joinedStanfordPresentFinal <- joinedStanfordPresentFinal %>%
  mutate(Party = Party.x)

joinedAll <- inner_join(massAll, totalCongressDistricts, by = c("City.Or.County" = "City", "State"))
joinedAllTemp <- left_join(joinedAll, dataCongress113, by = c("District", "State", "CongressNum"))
joinedAllFinal <- left_join(joinedAllTemp, dataCongress114, by = c("District", "State", "CongressNum"))
for (i in 1:893){
  if (is.na(joinedAllFinal$Party.x[i])) {
    joinedAllFinal$Party.x[i] <- joinedAllFinal$Party.y[i]
  }
}

joinedAllFinal <- joinedAllFinal %>%
  mutate(Party = Party.x)

joinedAllFinalv2 <- filter(joinedAllFinal, Party == "Democrat" | Party == "Republican")
```

Diagonostics

Our initial logistic regression model is based on political affiliation of the shooting location, week when the shooting occurred, number of victims killed, and number of victims injured had an AUC of .7375. We then tried a second model using only political affiliation and week. This model had an AUC of .7611 on the test set. An AUC of .5 means random predictive power, and an AUC of 1 means perfect predictive power. In this case, an AUC of 0.5 would mean that we could not predict whether a mass shooting is covered.



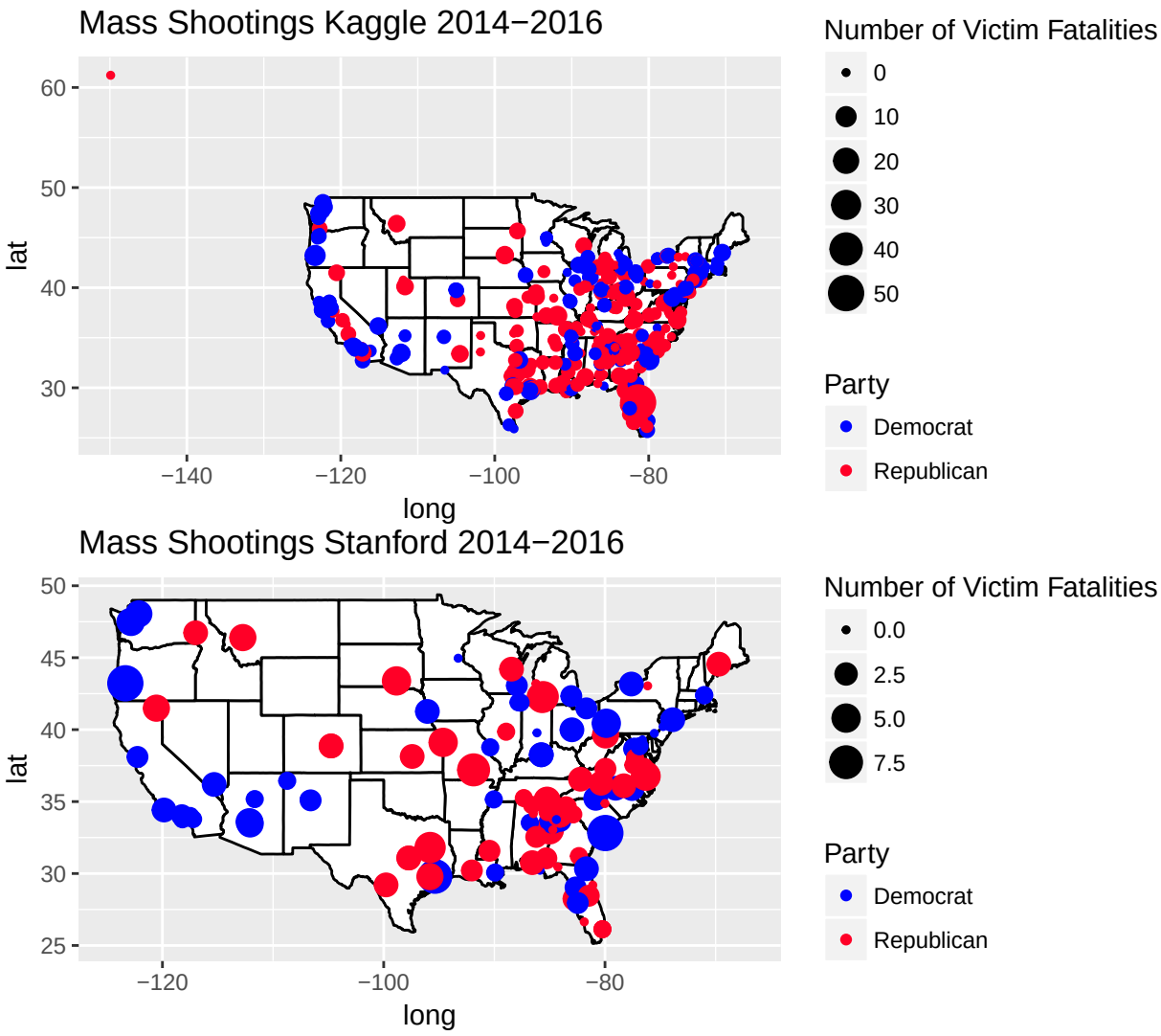

```
## Area under the curve: 0.7611
```



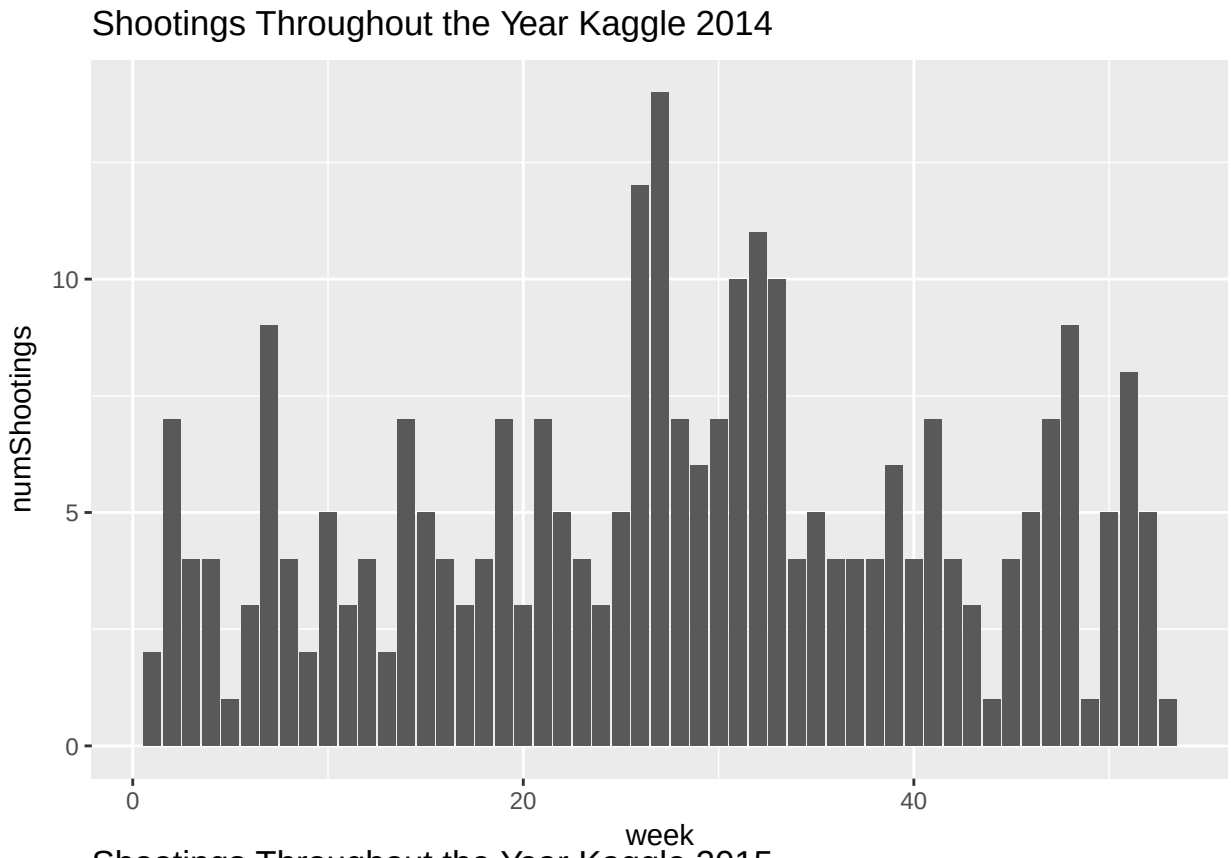

```
## Area under the curve: 0.7375
```

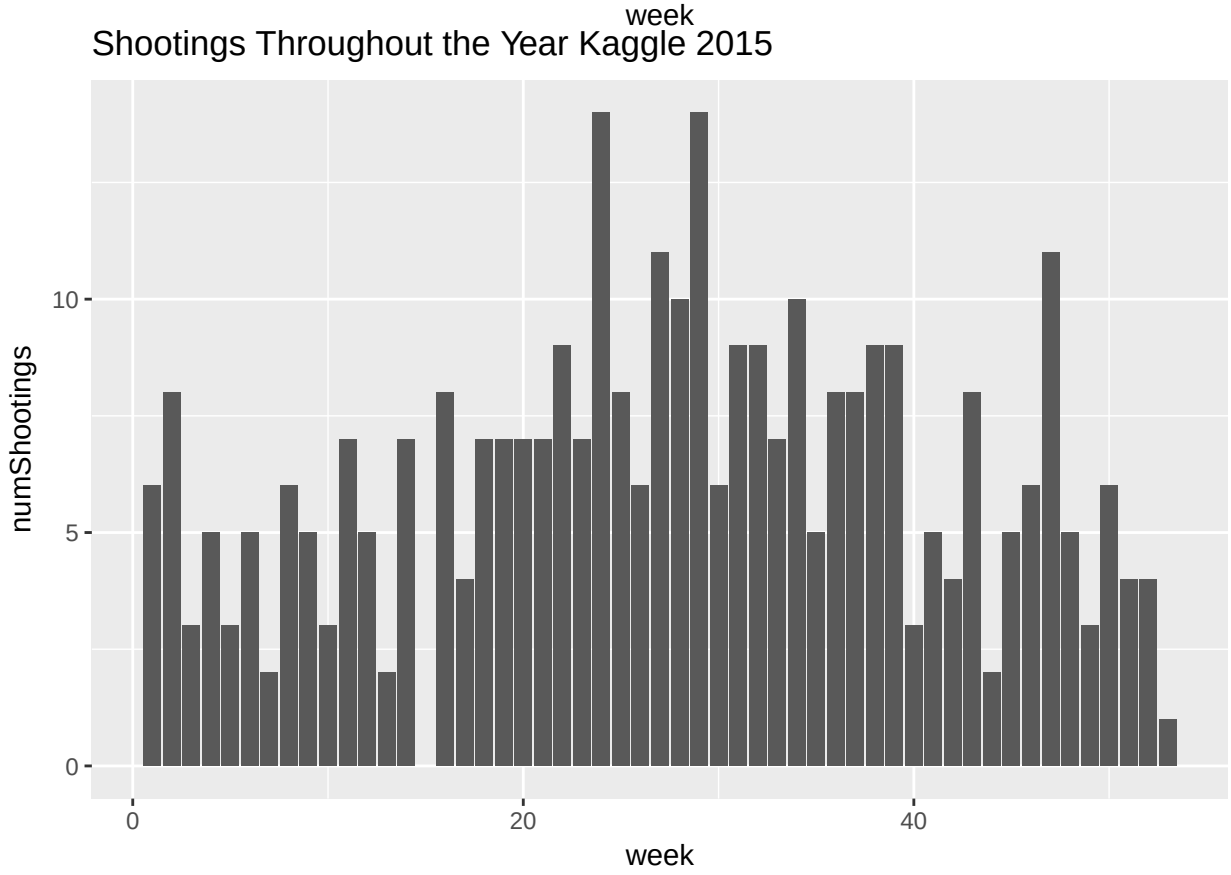
From the Kaggle Map, we can see that there are far more blue dots than red dots, meaning there were more mass shootings in democratic districts. The dots are also concentrated on the coasts, especially on cities and

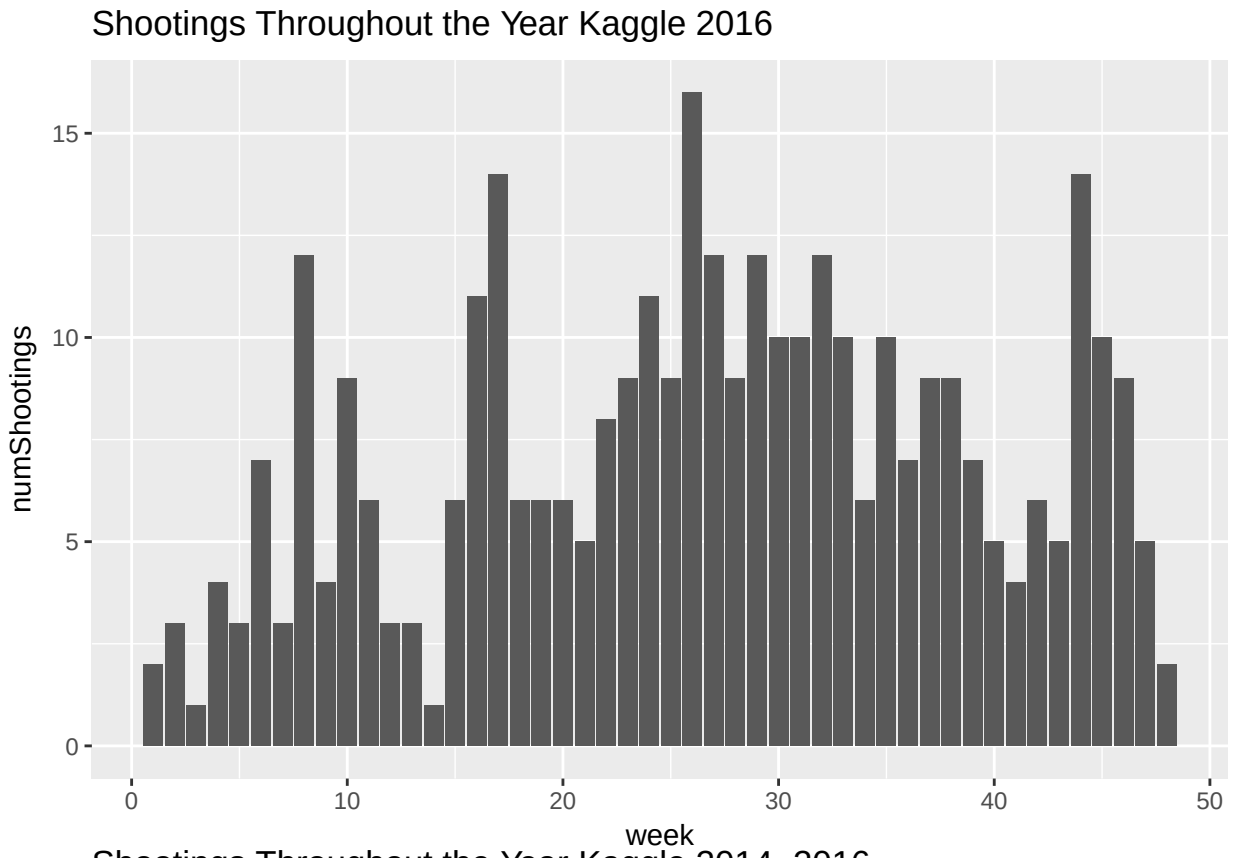
places with high population. Nearly all the shootings in the midwest were in republican districts. From the Stanford Map, there are significantly fewer dots in total than in the Kaggle Map. However, there appears to be a roughly similar proportion of red to blue dots.

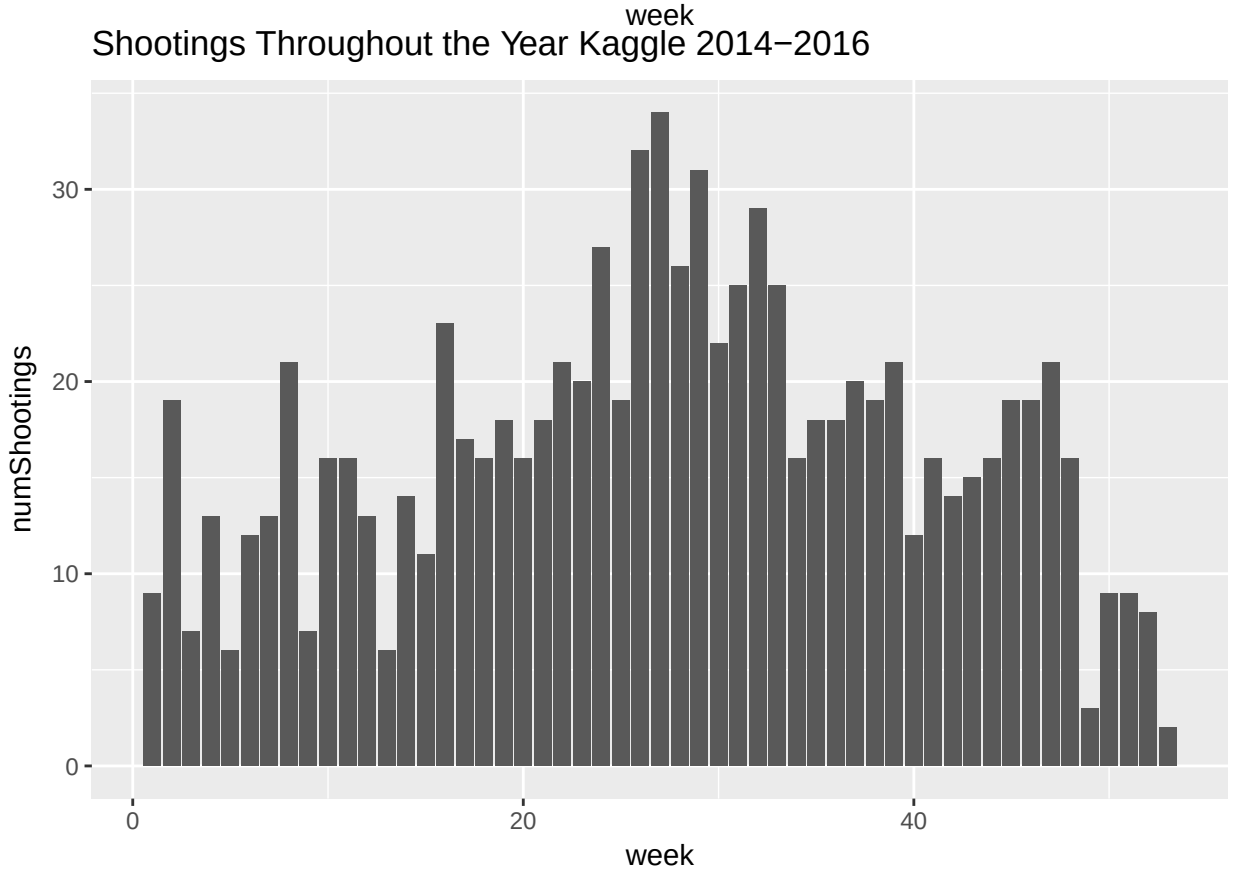


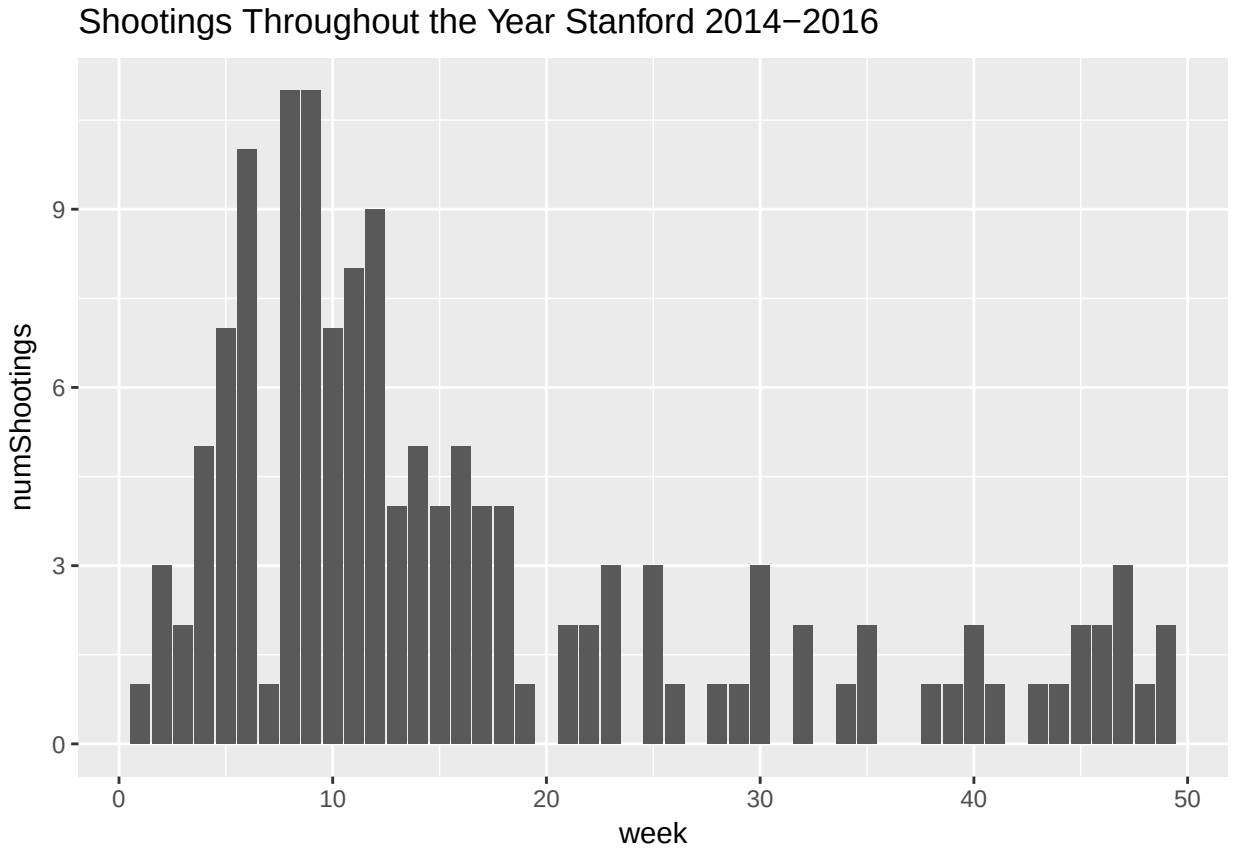
From the Kaggle Density plots, we can see that the distribution of shootings over the weeks of the year is approximately unimodal and approximately symmetric. Most shootings appear to consistently occur during the months of may, june, and july. Also to be noted is the scales on the axis, often there are more shootings than days in the week. However, when we look at the distribution of shootings throughout a year for the Stanford data, we see that it is noticeably skewed right. Here, mass shootings are more commonly reported in the months of January and February. The number of shootings per week is also far fewer.


Shootings Throughout the Year Kaggle 2014
numShootings
0
5
10
0
20
40
week


Shootings Throughout the Year Kaggle 2015
numShootings
0
5
10
0
20
40
week


Shootings Throughout the Year Kaggle 2016
numShootings
0
5
10
15
0
10
20
30
40
50
week


Shootings Throughout the Year Kaggle 2014−2016
numShootings
0
10
20
30
0
20
40
week



Results What we see is that the ratio of shootings in Republican to Democratic districts differs between the Stanford and Gun Violence Archive datasets. For the GVA data, the ratio we found was 0.433 while in the Stanford data, the ratio we found was 0.842. While we do not have enough information to tell us whether media sources overreport shootings in Republican districts or under report shootings in Democratic districts, we do have evidence that some sort of bias exists.

When we investigate when a shooting occurs during a year, we find more evidence suggesting a media reporting bias. For the years 2014 through 2016 in the GVA dataset, we can see that the distribution of shootings over the weeks of the year is approximately unimodal and approximately symmetric. However, when we look at the distribution of shootings throughout a year for the Stanford data, we see that it is noticeably skewed right. This difference suggests that media outlets tend to report more on shootings that occur near the start of a year and then drop off, even though shootings tend to mostly occur around the middle of a year. What this also does is confirm our suspicion of a "media fatigue" later in the year. Shootings are often reported by news outlets at the start of the year, but this reporting dies down as time goes on.

As discussed in the Diagnostics section, our initial logistic regression model is based on political affiliation of the shooting location, week when the shooting occurred, number of victims killed, and number of victims injured had an AUC of .7375. We then tried a second model using only political affiliation and week. This model had an AUC of .7611 on the test set. Since both of our models have a relatively high value, we appear to have some ability to predict whether or not a shooting would be reported by the media. We were surprised to see our model with only political and week performed better than our other model since we expected the severity of a shooting (number of victims killed and injured) to be a good predictor of the shooting receiving media coverage.

Conclusion We found evidence that suggests that media outlets have some bias in their reporting of mass shootings. At this time; however, we are not yet able to tell if media sources over report shootings in Republican districts or under report shootings in Democratic districts. We see some evidence of media fatigue from the differences in the density plots of shootings throughout the year. While we cannot definitively say

what explains these observations, the plots do display some sort of bias. Using the logistic model we created, we have some predictive power in determining whether or not a shooting is reported by the media. The model used party affiliation and week as variables, and had a AUC of .7611. Since this is significantly greater than 0.5 (random), party affiliation and week definitely played a part. We would like to improve our model, but this would require more data on mass shootings which we just do not have. Our observations are also limited to the near present (2014-2016) since the data we use comes from these three years. There is also an issue in how our two datasets define a mass shooting. Stanford defines it as three or more victims, while the GVA defines it as four or more victims. Furthermore, we do not have any way to predict when and where a shooting might occur. In the future, we would like to further explore the media's report of shootings. How do different news outlets report on mass shootings? Do these outlets' reporting differ depending on their political leaning?