Emily Mo - Final Project

Introduction

As someone who has never looked put much thought into the causes of arson before, yet is intrigued by social patterns, I wanted to investigate the demographics behind arson over the ten years of data in the NFIRS datasets. Therefore, I chose the focus question: Have any methods of arson or motivations behind arson grown significantly more or less prevalent over the past ten years? Does this imply anything about the changing state of American society?

Data Cleaning

I began by importing the arson.dbf tables from each year, which contained the basic information on each arson case. The Juvenile Subject tables contained valuable data on suspect demographics, which would be central to my project. First, I used the rbind function to combine each of the yearly tables for both the arson and the juvenile subject data.

```
f06 <- read.dbf("data/2006/NFIRS_2006_040108/arson.dbf")
f07 <- read.dbf("data/2007/NFIRS_2007_042309/arson.dbf")
f08 <- read.dbf("data/2008/NFIRS_2008_011910/arson.dbf")
f09 <- read.dbf("data/2009/NFIRS_2009_092710/arson.dbf")
f10 <- read.dbf("data/2010/NFIRS_2010_100711/arson.dbf")
f11 <- read.dbf("data/2011/NFIRS_2011_120612/arson.dbf")
f12 <- read.table("data/2012/NFIRS_2012_052714/arson.txt", sep = "^")
f13 <- read.table("data/2013/NFIRS_2013_121514/arson.txt", sep = "^")
f14 <- read.table("data/2014/NFIRS_2014_030216/arson.txt", sep = "^")
f15 <- read.table("data/2015/NFIRS_FIRES_2015_20170215/arson.txt",
    sep = "^")

j06 <- read.dbf("data/2006/NFIRS_2006_040108/arsonjuvsub.dbf")
j07 <- read.dbf("data/2007/NFIRS_2007_042309/arsonjuvsub.dbf")
j08 <- read.dbf("data/2008/NFIRS_2008_011910/arsonjuvsub.dbf")
j09 <- read.dbf("data/2009/NFIRS_2009_092710/arsonjuvsub.dbf")
j10 <- read.dbf("data/2010/NFIRS_2010_100711/arsonjuvsub.dbf")
j11 <- read.dbf("data/2011/NFIRS_2011_120612/arsonjuvsub.dbf")
j12 <- read.table("data/2012/NFIRS_2012_052714/arsonjuvsub.txt",
    sep = "^")
j13 <- read.table("data/2013/NFIRS_2013_121514/arsonjuvsub.txt",
    sep = "^")
j14 <- read.table("data/2014/NFIRS_2014_030216/arsonjuvsub.txt",
    sep = "^")
j15 <- read.table("data/2015/NFIRS_FIRES_2015_20170215/arsonjuvsub.txt",
    sep = "^")

colnames(f12) <- colnames(f06)
colnames(f13) <- colnames(f06)
colnames(f14) <- colnames(f06)
colnames(f15) <- colnames(f06)
dataf <- rbind(f06, f07, f08, f09, f10, f11, f12, f13, f14, f15)
```

```
colnames(j12) <- colnames(j06)
colnames(j13) <- colnames(j06)
colnames(j14) <- colnames(j06)
colnames(j15) <- colnames(j06)
dataj <- rbind(j06, j07, j08, j09, j10, j11, j12, j13, j14, j15)

dataf <- unique(dataf)
dataj <- unique(dataj)
```

Before joining the two new tables, I used the unique function to make sure there were no duplicate entries in either table.

Because the NFIRS Data Analysis Guide claims that each unique case is identifiable with "the combination of the State, fire department ID, incident date, incident number, and exposure number", I then joined the arson data with the juvenile subject data by those 5 variables so that no case would be over-counted.

```
data <- full_join(dataf, dataj, by = c("STATE", "FDID", "INC_DATE",
    "INC_NO", "EXP_NO"))
```

Then, because the dates in the table were in an unconvertible format, I created a loop to convert them to a convertible format so that the INC_DATE column could be converted to the date class.

```
for (i in 1:length(data$INC_DATE)) {
    if (nchar(as.character(data$INC_DATE[i])) == 7) {
        data$INC_DATE[i] <- paste("0", as.character(data$INC_DATE[i]),
            sep = "")
    }
}
```

I noticed that this loop took a very long time to run, and I wanted to access the results of this loop faster if I ever ran the code again, so I exported the resultant data to a CSV file so that I could quickly read it into the console in the future.

```
write.csv(data, file = "d.csv")
```

```
e <- read.csv("d.csv", sep = ",")
d <- e
```

The "d" dataset is now the source for all analysis in my project, though I still needed to clean the data further. Of course, I started with converting INC_DATE to the date class and creating a column called YEAR based off of INC_DATE so that I could more quickly analyze the data by year.

```
d$INC_DATE <- as.Date(d$INC_DATE, "%m%d%Y")
d$YEAR <- as.character(year(d$INC_DATE))
```

I then chose the variables that I wanted to work with, apart from the five main identifiers: Extent of fire, motivation factor, group involvement, ignition device, ignition fuel, gender, and race.

```
d <- select(d, STATE, FDID, INC_DATE, INC_NO, EXP_NO, SUB_SEQ_NO,
    EXT_FIRE, MOT_FACTS1, GRP_INVOL1, DEVI_IGNIT, DEVI_FUEL,
    GENDER, RACE, YEAR)
d$STATE <- as.character(d$STATE)
```

Next, I read in the code lookups so that I could replace the numeric codes for each variable with verbal labels.

```
c <- read.table("data/2015/NFIRS_FIRES_2015_20170215/codelookup.txt",
    sep = "^")
c$V1 <- as.character(c$V1)
```

```
DEVI_IGNIT <- subset(c, c$V1 == "DEVI_IGNIT")
DEVI_IGNIT <- DEVI_IGNIT[-1]
colnames(DEVI_IGNIT) <- c("DEVI_IGNIT", "ignitdef")
d <- left_join(d, DEVI_IGNIT, by = "DEVI_IGNIT")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

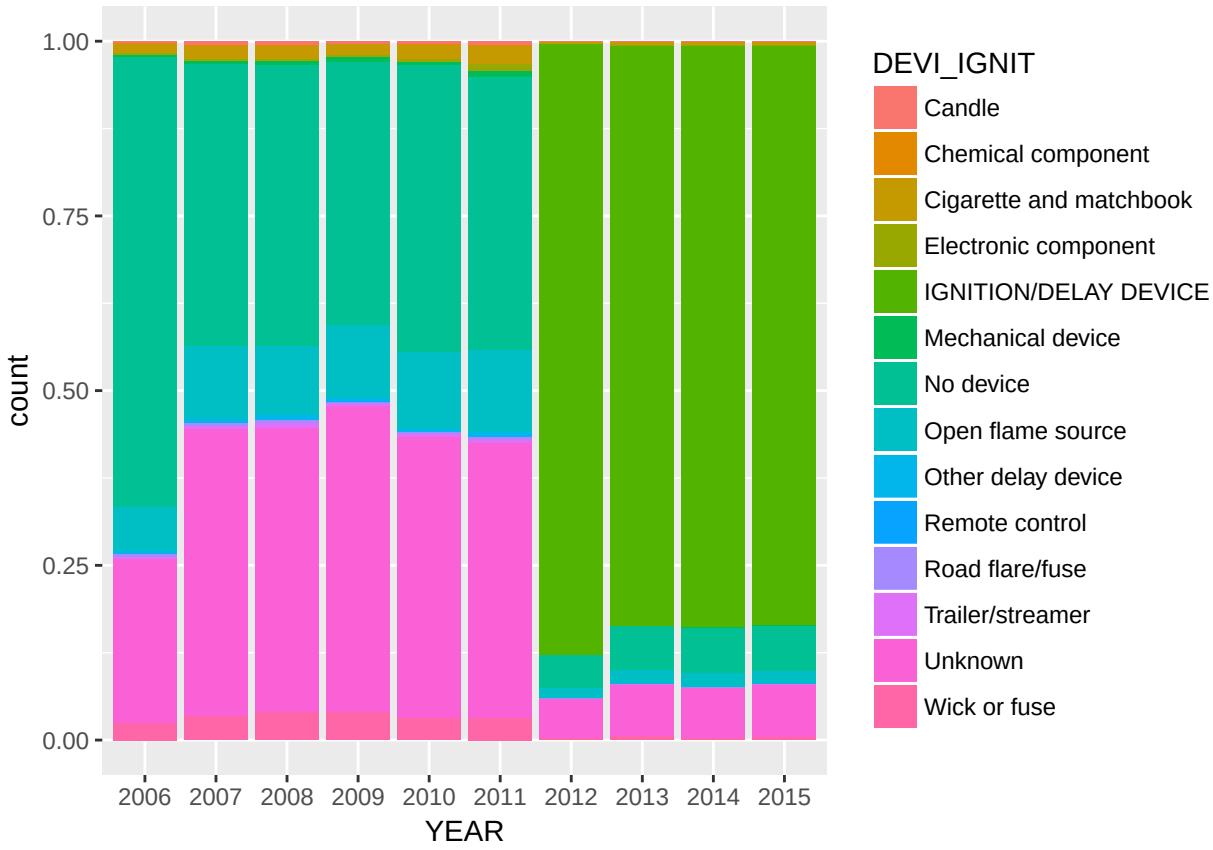
```
d <- select(d, -DEVI_IGNIT)
colnames(d)[colnames(d) == "ignitdef"] <- "DEVI_IGNIT"
d$DEVI_IGNIT <- as.character(d$DEVI_IGNIT)
```

When I made a table of each variable's numeric and verbal codes and joined it with the "d" table, then displayed all the levels of the variable, I found that each variable had a category that was named after the variable (for instance, DEVI_IGNIT had a level called "IGNITION/DELAY DEVICE", which doesn't actually provide any new information). When I graphed the proportions of ignition devices by year (graph shown below), it was clear that this miscellaneous "IGNITION/DELAY DEVICE" category only began to be used starting in 2012, so I combined this category with NA, as well as several other miscellaneous categories, as to focus more on the known data instead of the unknown.

```
levels(as.factor(d$DEVI_IGNIT))
```

```
##  [1] "Candle"                  "Chemical component"
##  [3] "Cigarette and matchbook" "Electronic component"
##  [5] "IGNITION/DELAY DEVICE"   "Mechanical device"
##  [7] "No device"               "Open flame source"
##  [9] "Other delay device"      "Remote control"
## [11] "Road flare/fuse"         "Trailer/streamer"
## [13] "Unknown"                 "Wick or fuse"
```

```
ggplot(data = subset(d, !is.na(DEVI_IGNIT)), aes(x = YEAR, fill = DEVI_IGNIT)) +
    geom_bar(position = "fill")  #
```




```
d$DEVI_IGNIT[d$DEVI_IGNIT == "NN"] <- NA
d$DEVI_IGNIT[d$DEVI_IGNIT == ""] <- NA
d$DEVI_IGNIT[d$DEVI_IGNIT == "DEVI_IGNIT"] <- NA
d$DEVI_IGNIT[d$DEVI_IGNIT == "IGNITION/DELAY DEVICE"] <- NA
dDEVI <- subset(d, !is.na(DEVI_IGNIT))
```

I also created a separate data table called dDEVI that contained all of the instances where DEVI_IGNIT was not NA, just to make analysis easier when examining that variable.

I repeated a similar process for the rest of the variables.

```
EXT_FIRE <- subset(c, c$V1 == "EXT_FIRE")
EXT_FIRE <- EXT_FIRE[-1]
colnames(EXT_FIRE) <- c("EXT_FIRE", "extdef")
EXT_FIRE$extdef <- paste(EXT_FIRE$EXT_FIRE, EXT_FIRE$extdef)
d <- left_join(d, EXT_FIRE, by = "EXT_FIRE")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
d <- select(d, -EXT_FIRE)
colnames(d)[colnames(d) == "extdef"] <- "EXT_FIRE"
levels(as.factor(d$EXT_FIRE))
```

```
## [1] " EXTENT OF FIRE"             "1 No flame or smoke showing"
## [3] "2 Smoke only showing"        "3 Flame and smoke showing"
## [5] "4 Fire through roof"         "5 Fully involved"
```

```
d$EXT_FIRE[d$EXT_FIRE == " EXTENT OF FIRE"] <- NA
dEXT <- subset(d, !is.na(EXT_FIRE))
```

```
MOT_FACTS1 <- subset(c, c$V1 == "MOT_FACTS1")
MOT_FACTS1 <- MOT_FACTS1[-1]
colnames(MOT_FACTS1) <- c("MOT_FACTS1", "def")
d <- left_join(d, MOT_FACTS1, by = "MOT_FACTS1")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
d <- select(d, -MOT_FACTS1)
colnames(d)[colnames(d) == "def"] <- "MOT_FACTS1"
d$MOT_FACTS1 <- as.character(d$MOT_FACTS1)
levels(as.factor(d$MOT_FACTS1))
```

```
##  [1] "Attention/sympathy"            "Auto theft concealment"
##  [3] "Burglary"                      "Burglary concealment"
##  [5] "Civil unrest"                  "Destroy records/evidence"
##  [7] "Domestic violence"             "Extortion"
##  [9] "Fireplay/curiosity"            "Foreclosed property"
## [11] "Hate crime"                    "Homicide"
## [13] "Homicide concealment"          "Institutional"
## [15] "Insurance fraud"               "Intimidation"
## [17] "Labor unrest"                  "Other motivation"
## [19] "Personal"                      "Protest"
## [21] "Sexual excitement"             "Societal"
## [23] "Suicide"                       "SUSPECTED MOTIVATION FACTORS 1"
## [25] "Thrills"                       "Unknown motivation"
## [27] "Vanity/recognition"            "Void contract/lease"
```

```
d$MOT_FACTS1[d$MOT_FACTS1 == "SUSPECTED MOTIVATION FACTORS 1"] <- NA
dMOT <- subset(d, !is.na(MOT_FACTS1))

RACE <- subset(c, c$V1 == "RACE")
RACE <- RACE[-1]
colnames(RACE) <- c("RACE", "def")
d <- left_join(d, RACE, by = "RACE")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
d <- select(d, -RACE)
colnames(d)[colnames(d) == "def"] <- "RACE"
d$RACE <- as.character(d$RACE)
levels(as.factor(d$RACE))
```

```
## [1] "American Indian, Eskimo or Aleut" "Asian"
## [3] "Black"                            "Other, includes multi-racial"
## [5] "RACE"                             "Undetermined"
## [7] "White"
```

```
d$RACE[d$RACE == "RACE"] <- NA
d$RACE[d$RACE == "U"] <- "Unknown"
dRAC <- subset(d, !is.na(RACE))

GRP_INVOL1 <- subset(c, c$V1 == "GRP_INVOL1")
GRP_INVOL1 <- GRP_INVOL1[-1]
colnames(GRP_INVOL1) <- c("GRP_INVOL1", "def")
```

```
d <- left_join(d, GRP_INVOL1, by = "GRP_INVOL1")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
d <- select(d, -GRP_INVOL1)
colnames(d)[colnames(d) == "def"] <- "GRP_INVOL1"
d$GRP_INVOL1 <- as.character(d$GRP_INVOL1)
levels(as.factor(d$GRP_INVOL1))
```

```
##  [1] "Anti-government group"
##  [2] "Gang"
##  [3] "GROUP INVOLVEMENT 1"
##  [4] "No group involvement, acted alone"
##  [5] "Organized crime"
##  [6] "Other group"
##  [7] "Outlaw motorcycle organization"
##  [8] "Racial/ethnic hate group"
##  [9] "Religious hate group"
## [10] "Sexual preference hate group"
## [11] "Terrorist group"
## [12] "Unknown"
```

```
d$GRP_INVOL1[d$GRP_INVOL1 == "GROUP INVOLVEMENT 1"] <- NA
d$GRP_INVOL1[d$GRP_INVOL1 == "U"] <- "Unknown"
dGRP <- subset(d, !is.na(d$GRP_INVOL1))

GENDER <- subset(c, c$V1 == "GENDER")
GENDER <- GENDER[-1]
colnames(GENDER) <- c("GENDER", "def")
d <- left_join(d, GENDER, by = "GENDER")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
d <- select(d, -GENDER)
colnames(d)[colnames(d) == "def"] <- "GENDER"
d$GENDER <- as.character(d$GENDER)
levels(as.factor(d$GENDER))
```

```
## [1] "Female" "Male"
```

```
dGEN <- subset(d, !is.na(GENDER))

DEVI_FUEL <- subset(c, c$V1 == "DEVI_FUEL")
DEVI_FUEL <- DEVI_FUEL[-1]
colnames(DEVI_FUEL) <- c("DEVI_FUEL", "def")
d <- left_join(d, DEVI_FUEL, by = "DEVI_FUEL")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
d <- select(d, -DEVI_FUEL)
colnames(d)[colnames(d) == "def"] <- "DEVI_FUEL"
d$DEVI_FUEL <- as.character(d$DEVI_FUEL)
levels(as.factor(d$DEVI_FUEL))
```

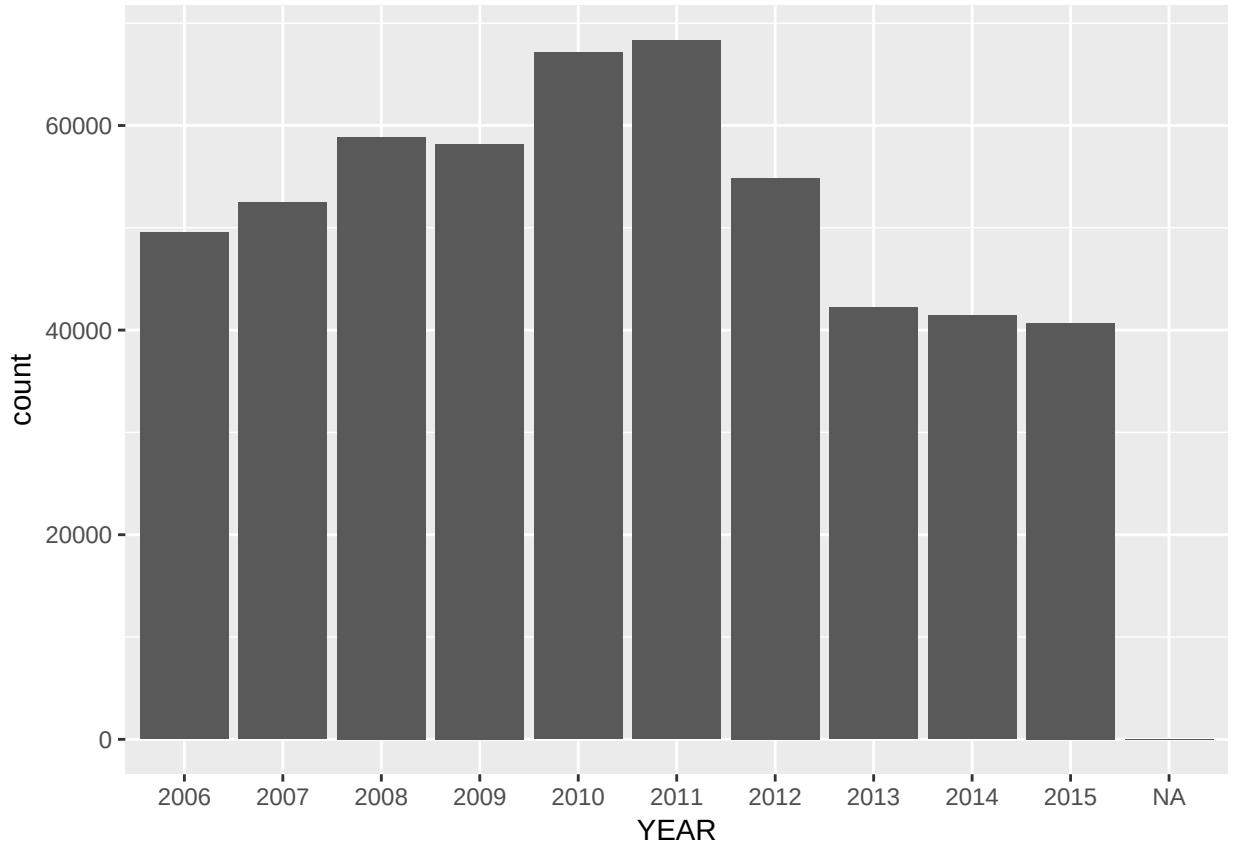
```
##  [1] "Explosive material"     "Flammable gas"
##  [3] "Ignitable liquid"       "Ignitable solid"
##  [5] "INCENDIARY DEVICE FUEL" "None"
##  [7] "Ordinary combustibles"  "Other material"
##  [9] "Pyrotechnic material"   "Unknown"
```

```
d$DEVI_FUEL[d$DEVI_FUEL == "INCENDIARY DEVICE FUEL"] <- NA
dFUE <- subset(d, !is.na(DEVI_FUEL))
```

Preliminary Graphics

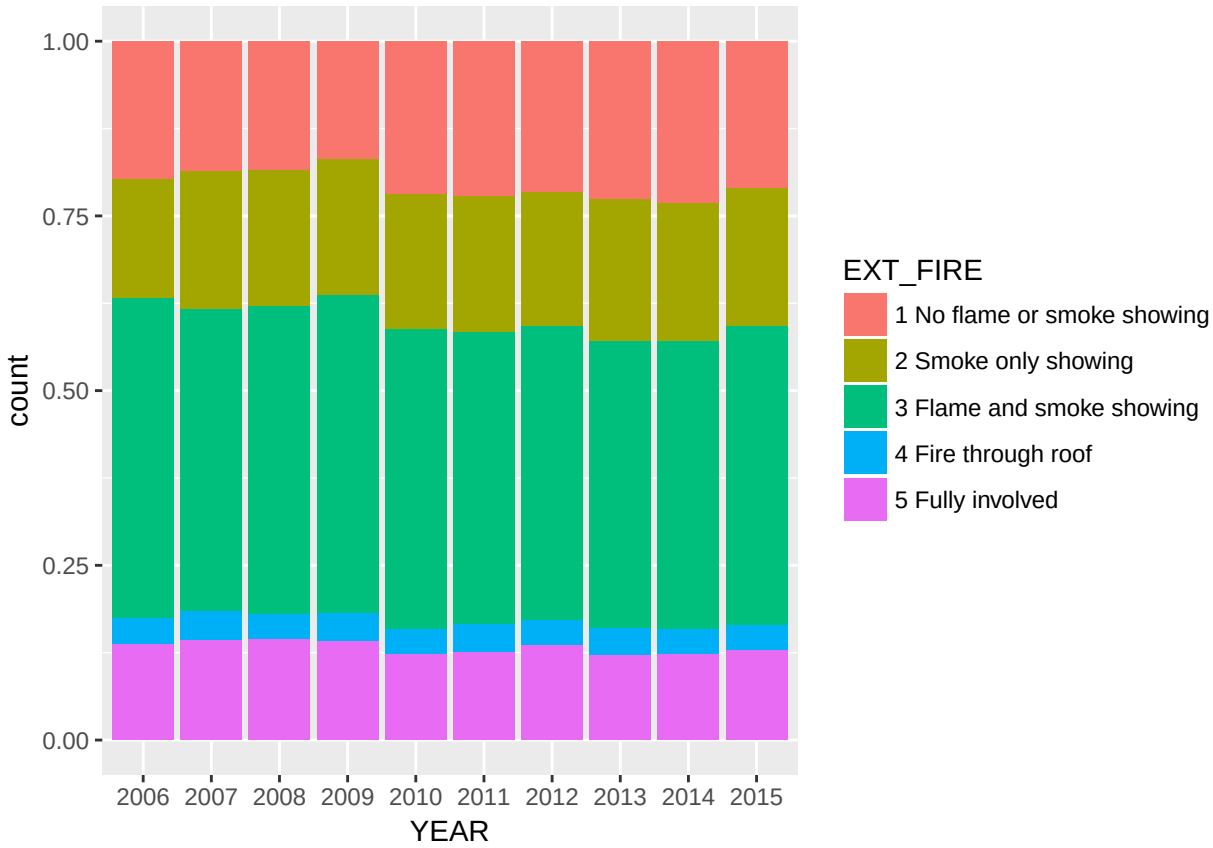
The following graphics were developed to help further visualize the data, detect any faults with the data, and provide ideas for my analysis.

```
ggplot(data = d, aes(x = YEAR)) + geom_bar()
```



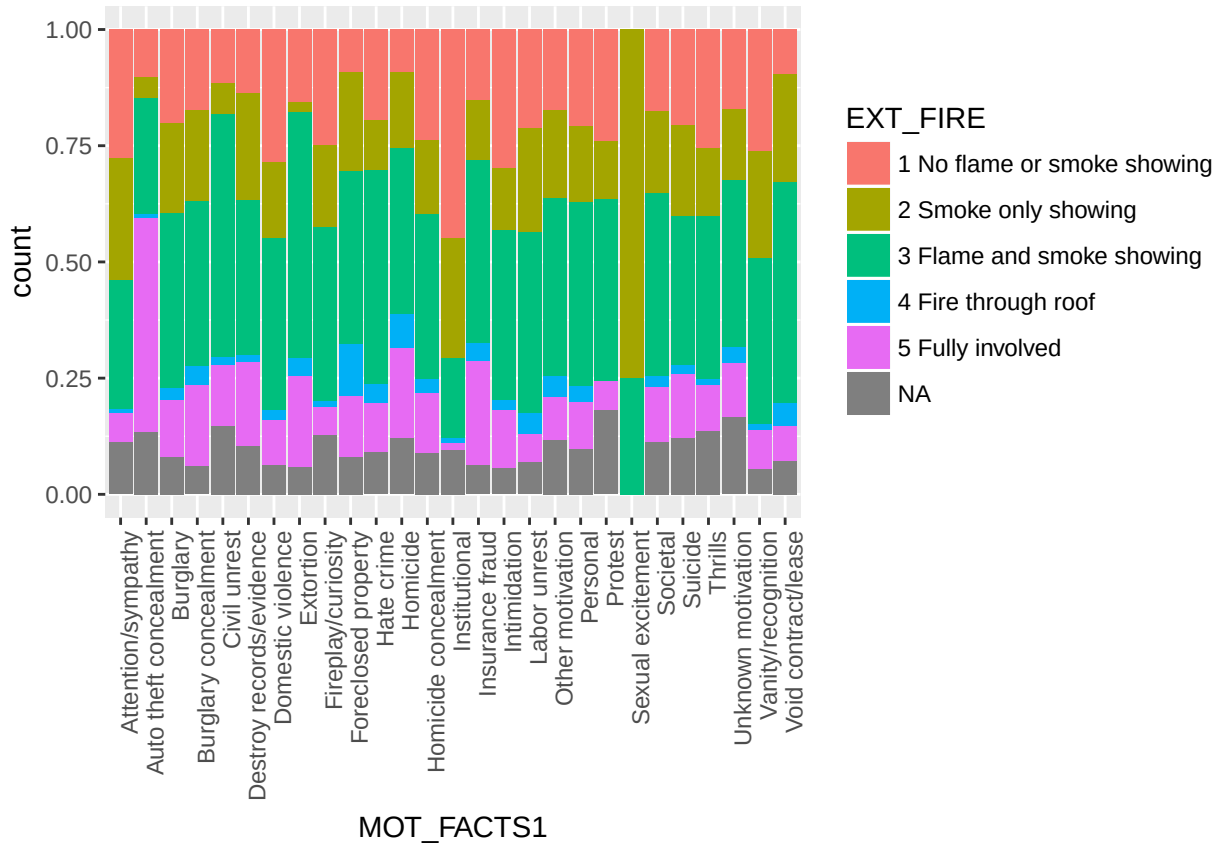
This graph shows the distribution of incidents over the years. Although there seems to be a sharp increase and then decrease around 2011, none of the years seem to have an alarmingly large or small number of incidents.

```
ggplot(data = dEXT, aes(x = YEAR, fill = EXT_FIRE)) + geom_bar(position = "fill")
```

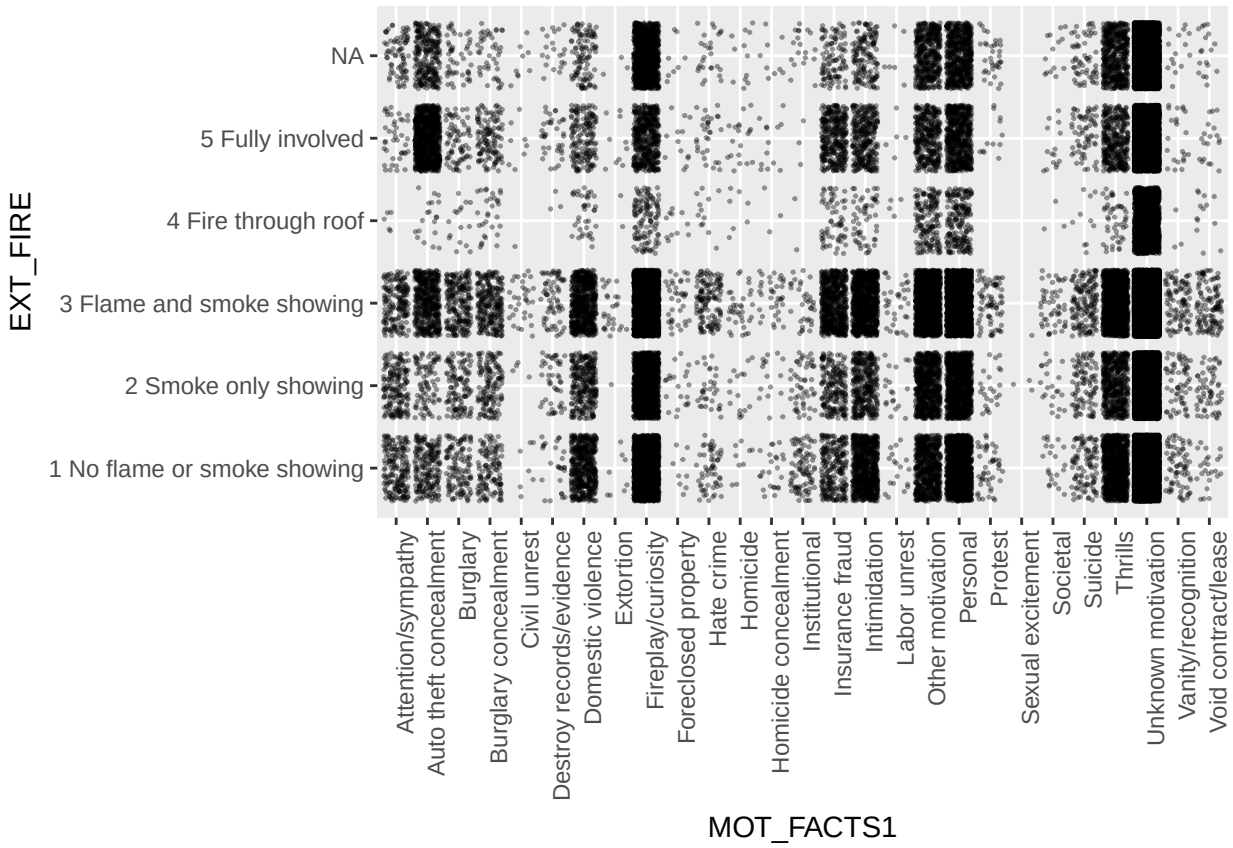


Here, I plotted the proportions of extent of fires caused by arson over the years. Again, from the visualization alone, no year seems to have an alarming proportion of a certain extent of fire, and the overall extent of fire does not seem to be getting worse or better with time.

```
ggplot(data = dMOT, aes(x = MOT_FACTS1, fill = EXT_FIRE)) + geom_bar(position = "fill") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

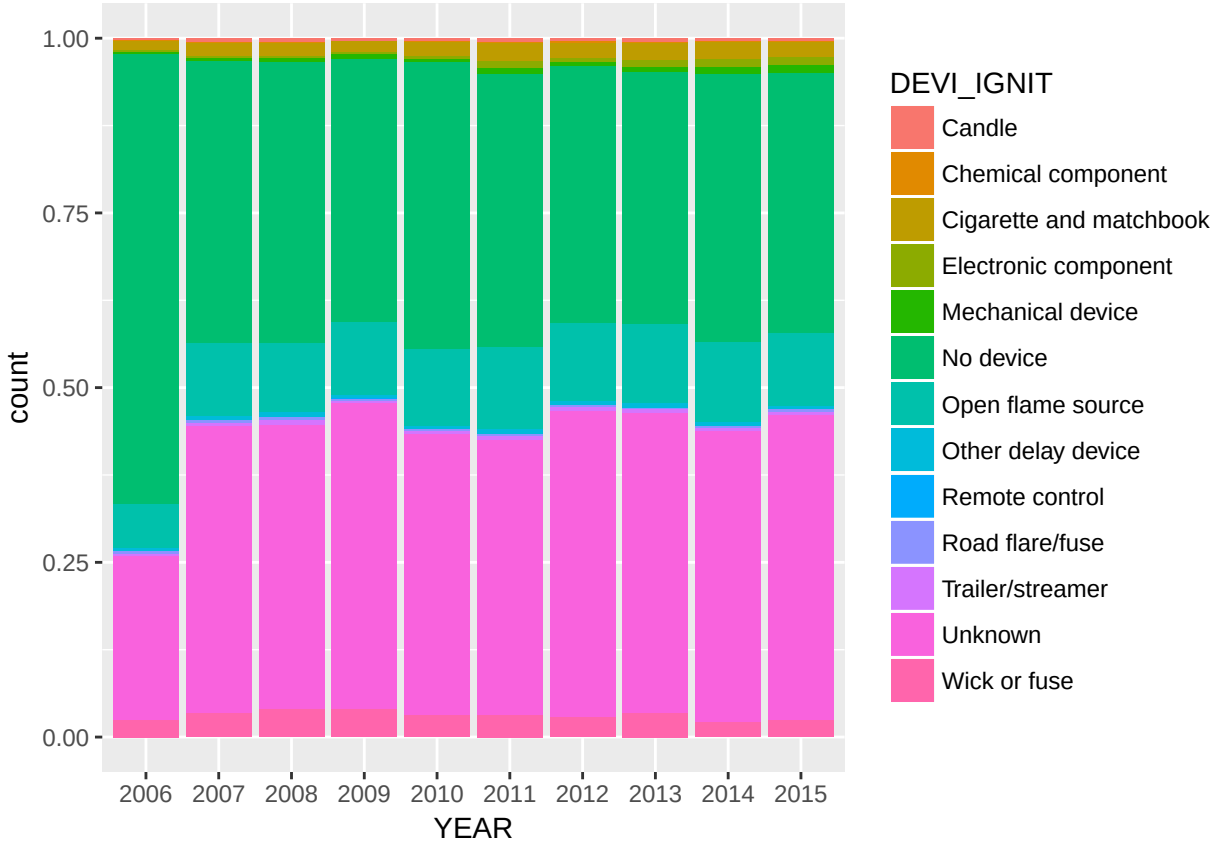



```
ggplot(data = dMOT, aes(x = MOT_FACTS1, y = EXT_FIRE)) + geom_jitter(size = 0.25,
    alpha = 0.4) + theme(axis.text.x = element_text(angle = 90,
    hjust = 1))
```



This graph was interesting because it showed clear variation of the extent of fire depending on the motivation behind the arson. I also made a jitter plot of the number of occurrences, rather than the proportion, of each extent of fire depending on motivation, because this helped to visualize the prominence of each motivation. I decided to run a general linear model on this relationship after seeing this variation.

```
ggplot(data = dDEVI, aes(x = YEAR, fill = DEVI_IGNIT)) + geom_bar(position = "fill")  #
```



Although I already graphed this relationship while cleaning the ignition device data, I examined it again after removing the miscellaneous categories. This revealed what seemed like a clear increase in mechanical and electronic ignition devices that I would later track the growth of by year.

I also wanted to see if I could detect any locational trends in the data, so I made a table with the latitude and longitude of all 50 states to left join to all of the data, as shown below.

```
STATE <- subset(c, c$V1 == "STATE")
STATE <- STATE[c(-1, -65, -66), ]
STATE <- select(STATE, V2)
STATE <- cbind(STATE, apply(STATE, 2, geocode))
STATE <- select(STATE, V2, V2.lon, V2.lat)
colnames(STATE) <- c("STATE", "state.lon", "state.lat")
STATE$STATE <- as.character(STATE$STATE)
```

Next, I made a table with the number of occurrences by state so that I could plot each of them on a map and visualize if there was a noticeable shift in locational arson incidents, or if there was an abnormally low or high number of arsons anywhere. To do this, I had to group the data by year and then sum the occurrences by year, then left join the state longitudes and latitudes to this yearly summary data.

```
nstate06 <- summarize(group_by(subset(d, d$YEAR == 2006), STATE),
    n())
nstate06 <- left_join(nstate06, STATE, by = "STATE")
nstate06 <- nstate06[-49, ]
colnames(nstate06)[2] <- "count"
nstate07 <- summarize(group_by(subset(d, d$YEAR == 2007), STATE),
    n())
nstate07 <- left_join(nstate07, STATE, by = "STATE")
nstate07 <- nstate07[1:47, ]
```

```
colnames(nstate07)[2] <- "count"
nstate08 <- summarize(group_by(subset(d, d$YEAR == 2008), STATE),
    n())
nstate08 <- left_join(nstate08, STATE, by = "STATE")
nstate08 <- nstate08[1:50, ]
colnames(nstate08)[2] <- "count"
nstate09 <- summarize(group_by(subset(d, d$YEAR == 2009), STATE),
    n())
nstate09 <- left_join(nstate09, STATE, by = "STATE")
nstate09 <- nstate09[1:51, ]
colnames(nstate09)[2] <- "count"
nstate10 <- summarize(group_by(subset(d, d$YEAR == 2010), STATE),
    n())
nstate10 <- left_join(nstate10, STATE, by = "STATE")
nstate10 <- nstate10[1:51, ]
colnames(nstate10)[2] <- "count"
nstate11 <- summarize(group_by(subset(d, d$YEAR == 2011), STATE),
    n())
nstate11 <- left_join(nstate11, STATE, by = "STATE")
nstate11 <- nstate11[1:51, ]
colnames(nstate11)[2] <- "count"
nstate12 <- summarize(group_by(subset(d, d$YEAR == 2012), STATE),
    n())
nstate12 <- left_join(nstate12, STATE, by = "STATE")
nstate12 <- nstate12[1:51, ]
colnames(nstate12)[2] <- "count"
nstate13 <- summarize(group_by(subset(d, d$YEAR == 2013), STATE),
    n())
nstate13 <- left_join(nstate13, STATE, by = "STATE")
nstate13 <- nstate13[1:50, ]
colnames(nstate13)[2] <- "count"
nstate14 <- summarize(group_by(subset(d, d$YEAR == 2014), STATE),
    n())
nstate14 <- left_join(nstate14, STATE, by = "STATE")
nstate14 <- nstate14[1:51, ]
colnames(nstate14)[2] <- "count"
nstate15 <- summarize(group_by(subset(d, d$YEAR == 2015), STATE),
    n())
nstate15 <- left_join(nstate15, STATE, by = "STATE")
nstate15 <- nstate15[1:51, ]
colnames(nstate15)[2] <- "count"
```

I retrieved my map by centering the map at Montana with a zoom level of 3, since this allowed for a view of all of the states. Then, I used ggplot2 to plot dots over each state indicating the number of arsons–the larger and the redder the dot, the more the occurrences in that state. The image below is my for 2006. To see maps for all years, see the interactive map feature in my webapp.

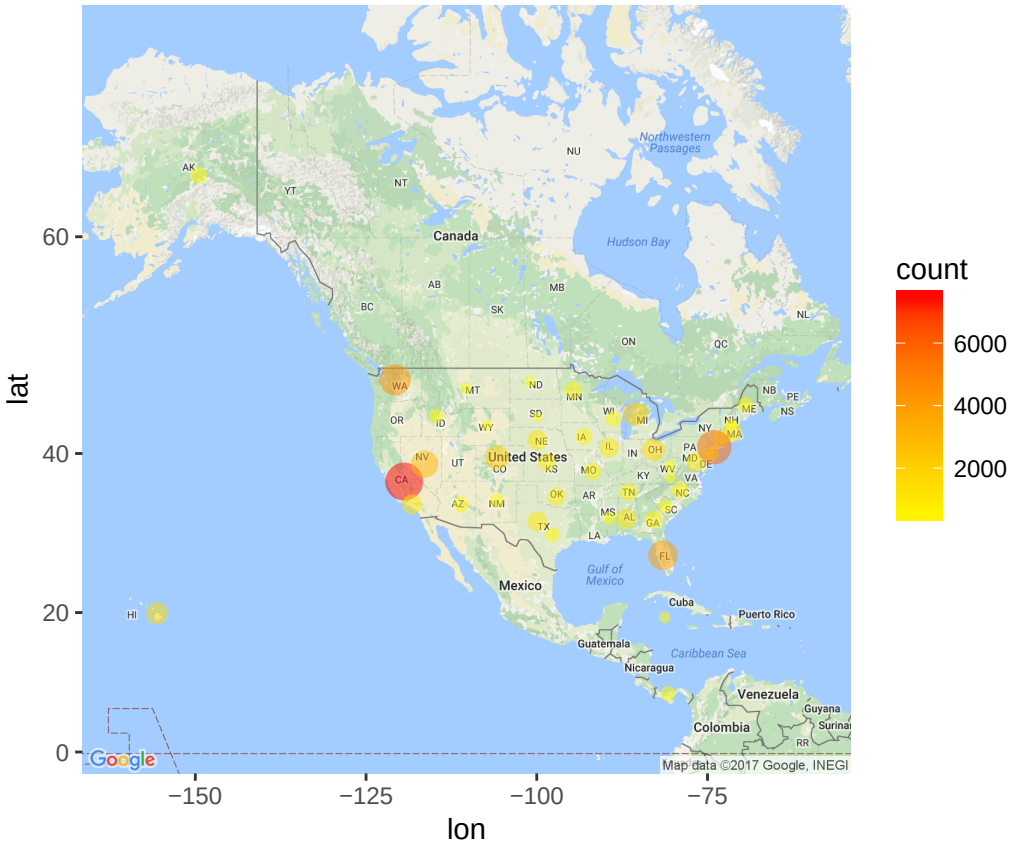
```
map1 <- get_map(location = "montana", source = "google", maptype = "roadmap",
    zoom = 3, scale = "auto")
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=montana&zoom=3&size=640x640&scal
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=montana&sensor=false
```

```
ggmap(map1) + geom_point(data = nstate06, aes(x = state.lon,
    y = state.lat, color = count, size = count), alpha = 0.5) +
    scale_size(guide = "none") + scale_color_gradient(low = "yellow",
    high = "red")
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



Trends in the Data

My main focus in my analysis was the proportions of the different types of ignition device, motivation, gender, race, group activity, and fuel in arsons over time. To perform this analysis, I had to code a "for" loop that would create a table of the partitions of each variable for every year.

1.) Ignition Device

For example, for ignition device, I first created a data frame of the proportions of each ignition device per year using the loop below:

```
dDEVIbyYR <- data.frame(matrix(ncol = 14))
colnames(dDEVIbyYR) <- c("Year", levels(as.factor(dDEVI$DEVI_IGNIT)))
for (year in 2006:2015) {
    vec <- c(year)
    total <- nrow(subset(d, !is.na(DEVI_IGNIT) & YEAR == year))
    for (i in 1:length(levels(as.factor(d$DEVI_IGNIT)))) {
        num <- nrow(subset(d, !is.na(DEVI_IGNIT) & YEAR == year &
            DEVI_IGNIT == levels(as.factor(DEVI_IGNIT))[i]))
```

```
        vec <- c(vec, num/total)
    }
    dDEVIbyYR <- rbind(dDEVIbyYR, vec)
}
dDEVIbyYR <- na.exclude(dDEVIbyYR)
```

The first three ignition device types of the resultant table are shown below:

	Year	Candle	Chemical component	Cigarette and matchbook
2	2006.00	0.00	0.00	0.01
3	2007.00	0.00	0.00	0.02
4	2008.00	0.00	0.00	0.02
5	2009.00	0.00	0.00	0.02
6	2010.00	0.00	0.00	0.02
7	2011.00	0.00	0.00	0.03
8	2012.00	0.00	0.00	0.02
9	2013.00	0.00	0.00	0.02
10	2014.00	0.00	0.00	0.02
11	2015.00	0.00	0.00	0.02

Prior to running any linear models on the data, I made a smooth plot of the proportions for each variable to determine whether a linear model would be appropriate. (Note that this was purely for visualization purposes, not to suggest that proportion vs. year could be treated as continuous.) I then ran a linear model for each type of ignition device. The two ignition device types who showed to have a significantly increasing or decreasing proportion over time were electronic components and mechanical devices, as summarized below.

```
lm1 <- lm(dDEVIbyYR$`Electronic component` ~ dDEVIbyYR$Year)
lm2 <- lm(dDEVIbyYR$`Mechanical device` ~ dDEVIbyYR$Year)
```

Proportion of arsons ignited with an electronic component vs. Year:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.2963	0.3131	-7.33	0.0001
dDEVIbyYR$Year	0.0011	0.0002	7.35	0.0001

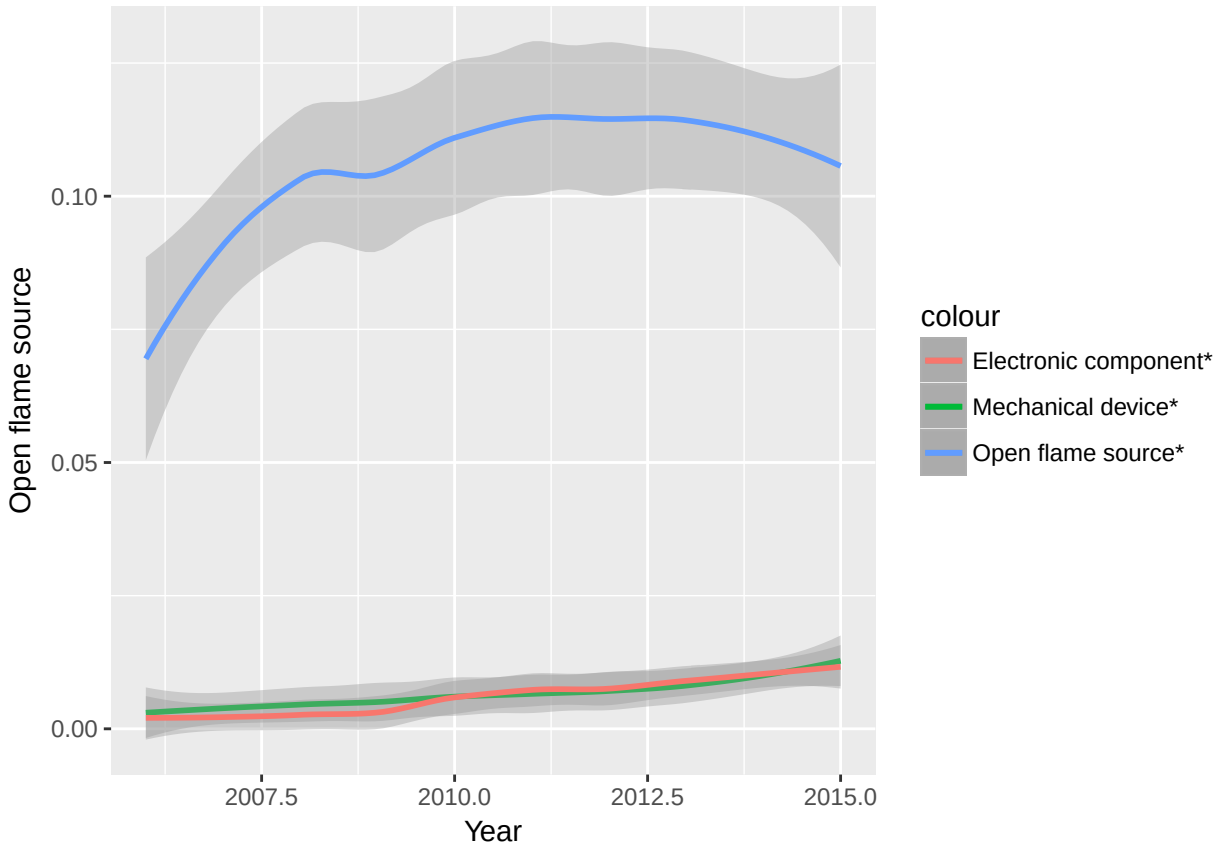
Proportion of arsons ignited with a mechanical device vs. Year:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.8648	0.3538	-5.27	0.0008
dDEVIbyYR$Year	0.0009	0.0002	5.29	0.0007

The linear trends for these significant relationships are depicted below:

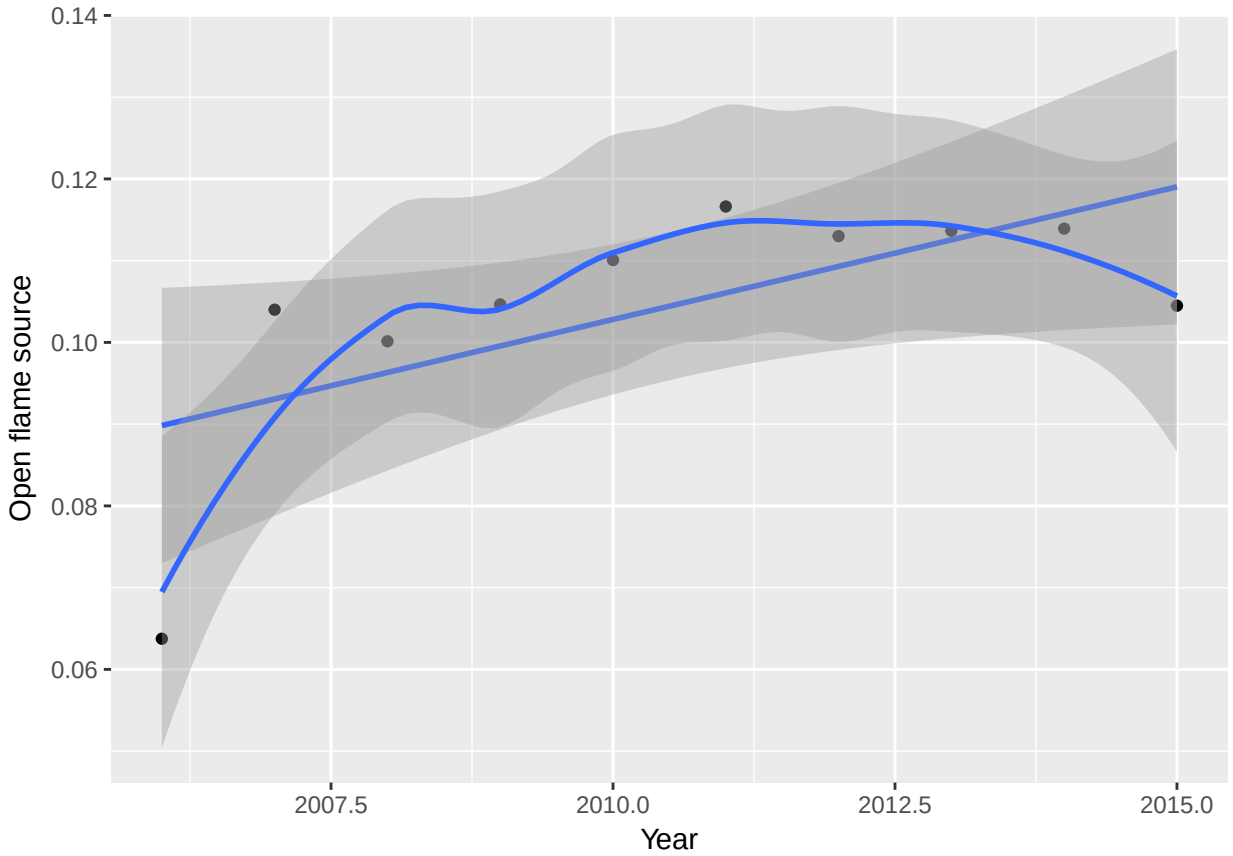
```
ggplot(data = dDEVIbyYR, aes(x = Year)) + stat_smooth(aes(y = `Open flame source`,
    col = "Open flame source*")) + stat_smooth(aes(y = `Mechanical device`,
    col = "Mechanical device*")) + stat_smooth(aes(y = `Electronic component`,
    col = "Electronic component*"))
```

```
## `geom_smooth()` using method = 'loess'
## `geom_smooth()` using method = 'loess'
## `geom_smooth()` using method = 'loess'
```



When examining the plot for the proportions of "Open flame source" over the years, it was clear that the relationship between proportion and year was more quadratic than linear. Therefore, I fit a quadratic model to the data instead.

```
ggplot(data = dDEVIbyYR, aes(x = Year)) + geom_point(aes(y = `Open flame source`)) +
    stat_smooth(aes(y = `Open flame source`), method = "lm") +
    stat_smooth(aes(y = `Open flame source`), method = "loess")
```



Proportion of arsons ignited with an open flame source vs. Year:

```
lm2b <- lm(dDEVIbyYR$`Open flame source` ~ -(dDEVIbyYR$Year)^2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.2963	0.3131	-7.33	0.0001
dDEVIbyYR$Year	0.0011	0.0002	7.35	0.0001

The three model summaries indicate that there is an extremely low probability that there is no significant change in the proportions of electronic component, mechanical device, or open flame source use to start arsons over the years 2006 to 2015. On average, arsons started by electronic components and mechanical devices have risen, further suggesting that technology has made it increasingly easier to start fires, or that people who wish to commit arson have grown increasingly more likely to resort to technology to aid them.

However, I decided to withhold from making a solid conclusion on the "Open flame source" ignition device type, because it seemed like a catch-all label for uncategorized open flame sources–this category could include a burner or a cigarette, for example, but there are already more specific categories for these ignition devices in use. Therefore, there is a possibility that the quadratic relationship between year and proportion of arsons filed under "Open flame source" could simply indicate that there was an increase and subsequent decrease in the use of the term as a catch-all. More information on the meaning of the term would is required in order to come to a more absolute conclusion.

2.) Motivation

First, I generated the table of proportions just as I did for ignition device:

```
dMOTbyYR <- data.frame(matrix(ncol = 28))
colnames(dMOTbyYR) <- c("Year", levels(as.factor(dMOT$MOT_FACTS1)))
for (year in 2006:2015) {
```

```
    vec <- c(year)
    total <- nrow(subset(d, !is.na(MOT_FACTS1) & YEAR == year))
    for (i in 1:length(levels(as.factor(d$MOT_FACTS1)))) {
        num <- nrow(subset(d, !is.na(MOT_FACTS1) & YEAR == year &
            MOT_FACTS1 == levels(as.factor(MOT_FACTS1))[i]))
        vec <- c(vec, num/total)
    }
    dMOTbyYR <- rbind(dMOTbyYR, vec)
}
dMOTbyYR <- na.exclude(dMOTbyYR)
```

I found only one motivation factor that changed significantly with the years, which was foreclosed property. The linear model showed that the proportion of arson motivated by foreclosed property has generally increased from 2006 to 2015, which is interesting because I predicted there to be a spike around the 2008 recession. However, there were actually no reports of arson motivated by foreclosed property until 2012. Therefore, looking at the graph of the proportion of foreclosure-motivated arsons vs. time, I decided that a linear model may not be entirely appropriate to explain this trend.

Proportion of arsons motivated by foreclosed property vs. Year:

```
lm3 <- lm(dMOTbyYR$`Foreclosed property` ~ dMOTbyYR$Year)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.0076	0.1892	-5.33	0.0007
dMOTbyYR$Year	0.0005	0.0001	5.33	0.0007

ggplot(data = dMOTbyYR, aes(x = Year, y = `Foreclosed property`)) + stat_smooth(col = "red")

3.) Gender

Generating the table of proportions:

```
dGENbyYR <- data.frame(matrix(ncol = 3))
colnames(dGENbyYR) <- c("Year", levels(as.factor(dGEN$GENDER)))
for (year in 2006:2015) {
    vec <- c(year)
    total <- nrow(subset(d, !is.na(GENDER) & YEAR == year))
    for (i in 1:length(levels(as.factor(d$GENDER)))) {
        num <- nrow(subset(d, !is.na(GENDER) & YEAR == year &
            GENDER == levels(as.factor(GENDER))[i]))
        vec <- c(vec, num/total)
    }
    dGENbyYR <- rbind(dGENbyYR, vec)
}
```

Because the gender variable had only two options, male of female, both genders showed significant, complimentary linear change in proportion over the years. Generally, the proportion of arsons committed by females have risen, and the proportion of arsons committed by males have fallen, though males remain the suspects of the vast majority of arsons. If I were to further look into this phenomenon, I would investigate the rising prominence of female arsons as a potential effect of the decreasing portrayals of women as submissive and well-behaved in the media.

```
lm4 <- lm(dGENbyYR$Male ~ dGENbyYR$Year)
summary(lm4)
```

```
##
## Call:
```

```
## lm(formula = dGENbyYR$Male ~ dGENbyYR$Year)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0108133 -0.0050708  0.0007738  0.0042468  0.0112955
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.3443786  1.6355901   8.159 3.79e-05 ***
## dGENbyYR$Year -0.0062628  0.0008135  -7.698 5.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007389 on 8 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8811, Adjusted R-squared:  0.8662
## F-statistic: 59.26 on 1 and 8 DF,  p-value: 5.752e-05
```

```
lm4a <- lm(dGENbyYR$Female ~ dGENbyYR$Year)
summary(lm4a)
```

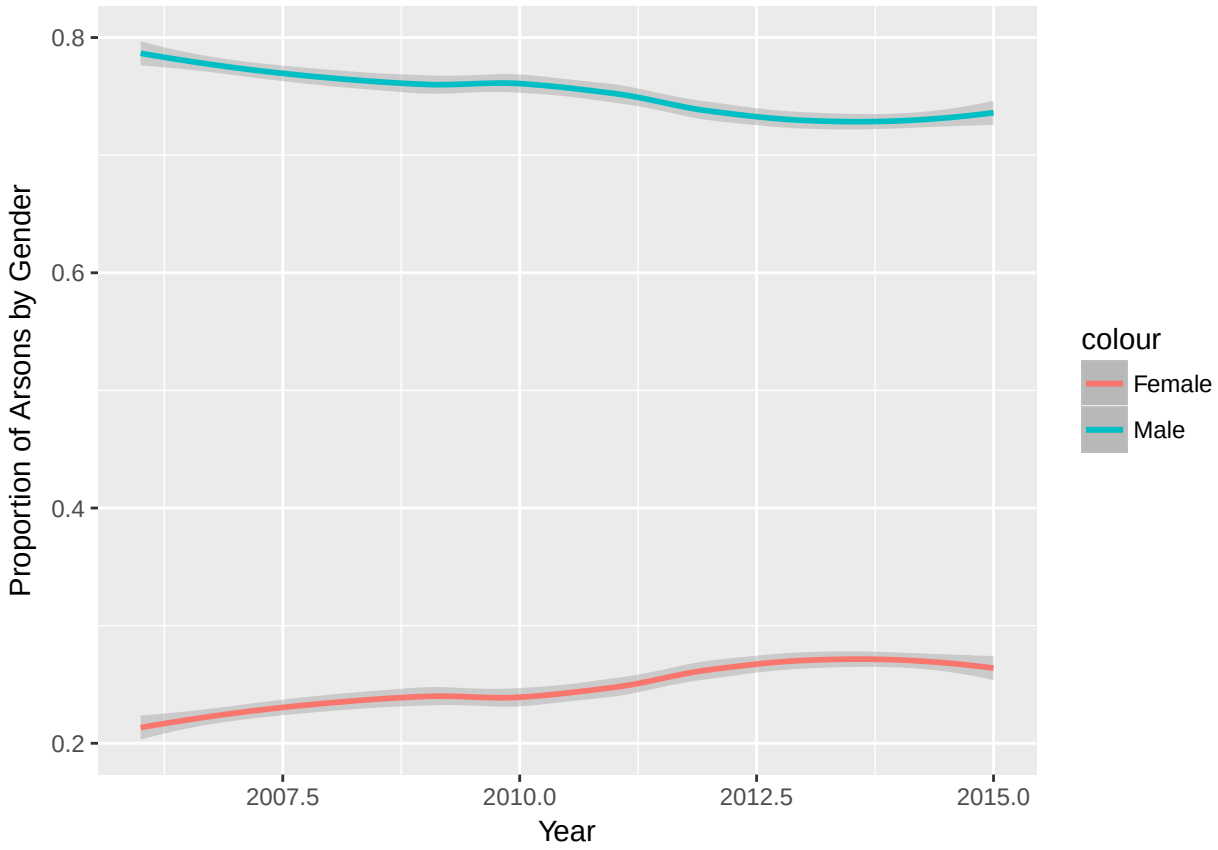
```
##
## Call:
## lm(formula = dGENbyYR$Female ~ dGENbyYR$Year)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0112955 -0.0042468 -0.0007738  0.0050708  0.0108133
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.234e+01  1.636e+00  -7.547 6.62e-05 ***
## dGENbyYR$Year  6.263e-03  8.135e-04   7.698 5.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007389 on 8 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8811, Adjusted R-squared:  0.8662
## F-statistic: 59.26 on 1 and 8 DF,  p-value: 5.752e-05
```

Proportion of arsons committed by males vs. Year:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.0076	0.1892	-5.33	0.0007
dMOTbyYR$Year	0.0005	0.0001	5.33	0.0007

```
ggplot(dGENbyYR, aes(Year)) + stat_smooth(aes(y = Male, colour = "Male")) +
    stat_smooth(aes(y = Female, colour = "Female")) + ylab("Proportion of Arsons by Gender")
```

```
## `geom_smooth()` using method = 'loess'
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
## `geom_smooth()` using method = 'loess'
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```



4.) Race

```
dRACbyYR <- data.frame(matrix(ncol = 7))
colnames(dRACbyYR) <- c("Year", levels(as.factor(dRAC$RACE)))
for (year in 2006:2015) {
    vec <- c(year)
    total <- nrow(subset(d, !is.na(RACE) & YEAR == year))
    for (i in 1:length(levels(as.factor(d$RACE)))) {
        num <- nrow(subset(d, !is.na(RACE) & YEAR == year & RACE ==
            levels(as.factor(RACE))[i]))
        vec <- c(vec, num/total)
    }
    dRACbyYR <- rbind(dRACbyYR, vec)
}
```

After running linear models of the proportions of all of the listed races vs. year, I found that there was a significant increase in proportion of arsons committed by black people, and a significant decrease in the proportion of arsons committed by people who were white, mixed race/non-listed races (non-listed means that they did not fit into any of the following categories: white, black, Asian, American Indian, Eskimo or Aleut, or Undetermined).

Proportion of arsons committed by white people vs. Year:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.5196	4.1566	4.70	0.0015
dRACbyYR$Year	-0.0094	0.0021	-4.55	0.0019

Proportion of arsons committed by black people vs. Year:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-16.8966	3.6923	-4.58	0.0018
dRACbyYR$Year	0.0085	0.0018	4.65	0.0016

Proportion of arsons committed by people of "other"/mixed race vs. Year:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.9628	1.9243	2.58	0.0327
dRACbyYR$Year	-0.0024	0.0010	-2.54	0.0348

5.) Group Involvement

Generating the table of proportions:

```
dGRPbyYR <- data.frame(matrix(ncol = 12))
colnames(dGRPbyYR) <- c("Year", levels(as.factor(dGRP$GRP_INVOL1)))
for (year in 2006:2015) {
    vec <- c(year)
    total <- nrow(subset(d, !is.na(GRP_INVOL1) & YEAR == year))
    for (i in 1:length(levels(as.factor(d$GRP_INVOL1)))) {
        num <- nrow(subset(d, !is.na(GRP_INVOL1) & YEAR == year &
            GRP_INVOL1 == levels(as.factor(GRP_INVOL1))[i]))
        vec <- c(vec, num/total)
    }
    dGRPbyYR <- rbind(dGRPbyYR, vec)
}
```

The linear models that I ran indicated significant linear relationships between three proportions of group involvement types vs. year, these types being "No group involvement" and "Gang". The linear models suggested that on average, the proportion of arsons committed by people who acted alone increased from 2006-2015, while the proportion of arsons committed by gangs decreased over the years. This phenomenon could be further investigated alongside the development of gang culture–are gangs becoming less prominent, and if so, are individuals performing the same criminal actions as gangs did, but alone?

Proportion of arsons committed with no group involvement vs. Year:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-33.8976	10.7183	-3.16	0.0133
dGRPbyYR$Year	0.0171	0.0053	3.20	0.0126

Proportion of arsons committed with no group involvement vs. Year:

6.) Fuel

Generating the table of proportions:

```
dGRPbyYR <- data.frame(matrix(ncol = 12))
colnames(dGRPbyYR) <- c("Year", levels(as.factor(dGRP$GRP_INVOL1)))
for (year in 2006:2015) {
    vec <- c(year)
    total <- nrow(subset(d, !is.na(GRP_INVOL1) & YEAR == year))
    for (i in 1:length(levels(as.factor(d$GRP_INVOL1)))) {
        num <- nrow(subset(d, !is.na(GRP_INVOL1) & YEAR == year &
            GRP_INVOL1 == levels(as.factor(GRP_INVOL1))[i]))
        vec <- c(vec, num/total)
    }
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0811	0.8492	2.45	0.0399
dGRPbyYR$Year	-0.0010	0.0004	-2.44	0.0407

```
    dGRPbyYR <- rbind(dGRPbyYR, vec)
}
```

The linear models that I ran indicated significant linear relationships between four proportions of fuel types used in arson vs. year, these types being "Pyrotechnic material", "flammable gas", "ignitable liquid", and "unknown". The models suggested that the proportions of pyrotechnic material, flammable gas, and ignitable liquid all generally decreased over the years of 2006-2015, while the proportion of unknown fuel type generally rose. According to the data analysis guide provided by NFIRS, "null and blank values are considered unreported data and differ in meaning and substance from 'unknown' data" (pg. 12) because "unknown" data indicates a genuine lack of knowledge, so this linear trend potentially indicates that arsonists are getting better at concealing their means of arson.

Proportions of arsons committed with pyrotechnic material vs. year

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.3420	0.5677	4.13	0.0033
dFUEbyYR$Year	-0.0012	0.0003	-4.11	0.0034

Proportions of arsons committed with flammable gas vs. year

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5635	0.7979	3.21	0.0124
dFUEbyYR$Year	-0.0013	0.0004	-3.16	0.0135

Proportions of arsons committed with ignitable liquid vs. year

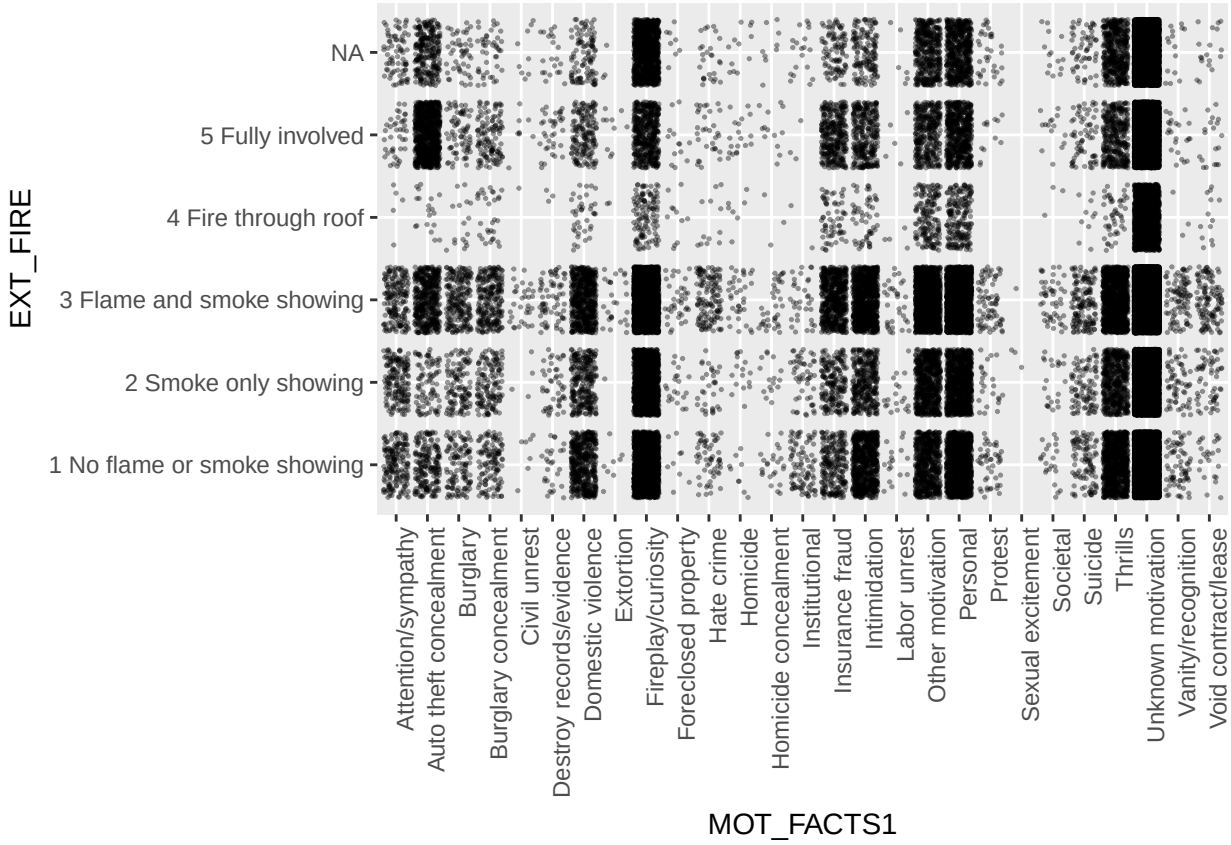
Proportions of arsons committed with unknown fuel vs. year

Motivation Factor vs. Extent of Fire

```
ggplot(data = dMOT, aes(x = MOT_FACTS1, y = EXT_FIRE)) + geom_jitter(size = 0.25,
    alpha = 0.4) + theme(axis.text.x = element_text(angle = 90,
    hjust = 1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0890	1.5770	4.50	0.0020
dFUEbyYR$Year	-0.0034	0.0008	-4.39	0.0023

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.3738	1.9809	-3.22	0.0123
dFUEbyYR$Year	0.0033	0.0010	3.38	0.0096



Because the graphic I made of extent of fire vs. motivation factor showed obvious differences in the extent of fire depending on motivation factor, I wanted to run a general linear regression model on all the data to see if any of the motivations could serve as good predictors of the extent of the fire.

```
s4a <- subset(d, !is.na(MOT_FACTS1) & !is.na(EXT_FIRE))
m4a <- glm(as.factor(EXT_FIRE) ~ as.factor(MOT_FACTS1), family = binomial,
    data = s4a)
```

With an alpha of 0.05, the model suggests that arsons motivated by auto theft concealment, burglary, burglary concealment, civil unrest, destroying records/evidence, extortion, foreclosed property, hate crime, homicide, institutional reasons, insurance fraud, personal reasons, societal reasons, and suicide can be significant predictors of fire extent. Interestingly, out of these significant factors, foreclosed property has the most positive relationship with extent of fire, meaning that the extent of fire is expected to be highest for an arson motivated by foreclosed property. Overall, the data illustrates an interesting image about foreclosed property and arson: the proportion of arsons motivated by foreclosed property has generally risen through 2006-2015, and arsons caused by foreclosed property tend to be more consummate.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7870	0.0892	8.83	0.0000
as.factor(MOT_FACTS1)Auto theft concealment	1.2194	0.1171	10.41	0.0000
as.factor(MOT_FACTS1)Burglary	0.4893	0.1364	3.59	0.0003
as.factor(MOT_FACTS1)Burglary concealment	0.7043	0.1341	5.25	0.0000
as.factor(MOT_FACTS1)Civil unrest	1.0738	0.4160	2.58	0.0098
as.factor(MOT_FACTS1)Destroy records/evidence	0.9345	0.2272	4.11	0.0000
as.factor(MOT_FACTS1)Domestic violence	0.0361	0.1046	0.35	0.7300
as.factor(MOT_FACTS1)Extortion	0.8225	0.3974	2.07	0.0385
as.factor(MOT_FACTS1)Fireplay/curiosity	0.1295	0.0929	1.39	0.1634
as.factor(MOT_FACTS1)Foreclosed property	1.4225	0.3623	3.93	0.0001
as.factor(MOT_FACTS1)Hate crime	0.5171	0.1739	2.97	0.0030
as.factor(MOT_FACTS1)Homicide	1.3596	0.3634	3.74	0.0002
as.factor(MOT_FACTS1)Homicide concealment	0.2545	0.2536	1.00	0.3157
as.factor(MOT_FACTS1)Institutional	-0.7758	0.1741	-4.46	0.0000
as.factor(MOT_FACTS1)Insurance fraud	0.8568	0.1141	7.51	0.0000
as.factor(MOT_FACTS1)Intimidation	-0.0110	0.0996	-0.11	0.9123
as.factor(MOT_FACTS1)Labor unrest	0.4335	0.2827	1.53	0.1251
as.factor(MOT_FACTS1)Other motivation	0.6285	0.0997	6.30	0.0000
as.factor(MOT_FACTS1)Personal	0.4266	0.0949	4.49	0.0000
as.factor(MOT_FACTS1)Protest	0.0939	0.1967	0.48	0.6331
as.factor(MOT_FACTS1)Sexual excitement	9.7791	59.7341	0.16	0.8700
as.factor(MOT_FACTS1)Societal	0.6106	0.2542	2.40	0.0163
as.factor(MOT_FACTS1)Suicide	0.3956	0.1454	2.72	0.0065
as.factor(MOT_FACTS1)Thrills	0.0828	0.0975	0.85	0.3958
as.factor(MOT_FACTS1)Unknown motivation	0.5594	0.0900	6.22	0.0000
as.factor(MOT_FACTS1)Vanity/recognition	0.1707	0.1657	1.03	0.3028
as.factor(MOT_FACTS1)Void contract/lease	1.3868	0.2331	5.95	0.0000

Summary:

Given my knowledge and skill level with R and the the fact that I did not work on this project over a long period of time, I could not come to any astounding conclusions on arson and society–however, the linear trends that I observed of the proportions of arson characteristics over time can serve as a launching pad for an abundance of investigations like the ones I mentioned with the individual linear models above. Much of this project consisted of experimenting with ways to visualize the data, and most of those visualizations did not translate to further analysis, but it served as great practice in discovering a focus question through visuals. Of course, many statistically significant patterns are not as easily detectable, but the method I used worked well given the time that I had. I also felt that my very rudimentary knowledge of statistical methods combined with my lack of experience analyzing data with R limited the scope of my analysis, but I still did vastly expand my R arsenal simply through curiosity and trial and error.