Wild Fire Analysis

Introduction

This database is a spectacular collection of data on wildfires in the United States from 1992 to 2015 created to support the US Fire Program Analysis. It has data on nearly 2 million wildfires over this time period.

- Load Data
- Wildfires over Time
- Fires by Size
- Wildfire Causes
- Wildfires by Geography
- Target Feature analysys To get started, load the libraries that we will need. We'll want RSQLite and dbplyr to extract the data from the sqlite database. We want dplyr for manipulation and ggplot2 for plotting of course.

```
library(RSQLite)
library(dbplyr)
library(dplyr)
library(purrr)
library(ggplot2)
library(xts)
library(ggfortify)
library(ggthemes)
library(maps)
library(mapdata)
library(leaflet)
```

Load the Data

Let's get the data from the database. Because it will fit into RAM, we'll want to extract the data into a dataframe rather than running sql queries againt the database on disk because it will be faster.

```
# Create a db connection
connect <- dbConnect(SQLite(), '~/Downloads/FPA_FOD_20170508.sqlite')

# pull the fires table into RAM

fires <- tbl(connect,"Fires") %>% collect()

# check the size of the table
print(object.size(fires),units = "Gb")
```

```
## 0.9 Gb
```

```
# Disconnect from Db

dbDisconnect(connect)
```

Get a quick view of the data

```
glimpse(fires)
```

```
## Observations: 1,880,465
## Variables: 39
## $ OBJECTID                  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, ...
## $ FOD_ID                    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, ...
## $ FPA_ID                    <chr> "FS-1418826", "FS-1418827", "FS-141...
## $ SOURCE_SYSTEM_TYPE        <chr> "FED", "FED", "FED", "FED", "FED", ...
## $ SOURCE_SYSTEM             <chr> "FS-FIRESTAT", "FS-FIRESTAT", "FS-F...
## $ NWCG_REPORTING_AGENCY     <chr> "FS", "FS", "FS", "FS", "FS", "FS",...
## $ NWCG_REPORTING_UNIT_ID    <chr> "USCAPNF", "USCAENF", "USCAENF", "U...
## $ NWCG_REPORTING_UNIT_NAME  <chr> "Plumas National Forest", "Eldorado...
## $ SOURCE_REPORTING_UNIT     <chr> "0511", "0503", "0503", "0503", "05...
## $ SOURCE_REPORTING_UNIT_NAME <chr> "Plumas National Forest", "Eldorado...
## $ LOCAL_FIRE_REPORT_ID      <chr> "1", "13", "27", "43", "44", "54", ...
## $ LOCAL_INCIDENT_ID         <chr> "PNF-47", "13", "021", "6", "7", "8...
## $ FIRE_CODE                 <chr> "BJ8K", "AAC0", "A32W", NA, NA, NA,...
## $ FIRE_NAME                 <chr> "FOUNTAIN", "PIGEON", "SLACK", "DEE...
## $ ICS_209_INCIDENT_NUMBER   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ ICS_209_NAME              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ MTBS_ID                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ MTBS_FIRE_NAME            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ COMPLEX_NAME              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ FIRE_YEAR                 <int> 2005, 2004, 2004, 2004, 2004, 2004,...
## $ DISCOVERY_DATE            <dbl> 2453404, 2453138, 2453156, 2453184,...
## $ DISCOVERY_DOY             <int> 33, 133, 152, 180, 180, 182, 183, 6...
## $ DISCOVERY_TIME            <chr> "1300", "0845", "1921", "1600", "16...
## $ STAT_CAUSE_CODE           <dbl> 9, 1, 5, 1, 1, 1, 1, 5, 5, 1, 1, 1,...
## $ STAT_CAUSE_DESCR          <chr> "Miscellaneous", "Lightning", "Debr...
## $ CONT_DATE                 <dbl> 2453404, 2453138, 2453156, 2453190,...
## $ CONT_DOY                  <int> 33, 133, 152, 185, 185, 183, 184, 6...
## $ CONT_TIME                 <chr> "1730", "1530", "2024", "1400", "12...
## $ FIRE_SIZE                 <dbl> 0.10, 0.25, 0.10, 0.10, 0.10, 0.10,...
## $ FIRE_SIZE_CLASS           <chr> "A", "A", "A", "A", "A", "A", "A", ...
## $ LATITUDE                  <dbl> 40.03694, 38.93306, 38.98417, 38.55...
## $ LONGITUDE                 <dbl> -121.0058, -120.4044, -120.7356, -1...
## $ OWNER_CODE                <dbl> 5, 5, 13, 5, 5, 5, 5, 13, 13, 5, 5,...
## $ OWNER_DESCR               <chr> "USFS", "USFS", "STATE OR PRIVATE",...
## $ STATE                     <chr> "CA", "CA", "CA", "CA", "CA", "CA",...
## $ COUNTY                    <chr> "63", "61", "17", "3", "3", "5", "1...
## $ FIPS_CODE                 <chr> "063", "061", "017", "003", "003", ...
## $ FIPS_NAME                 <chr> "Plumas", "Placer", "El Dorado", "A...
## $ Shape                     <blob> blob[60 B], blob[60 B], blob[60 B]...
```

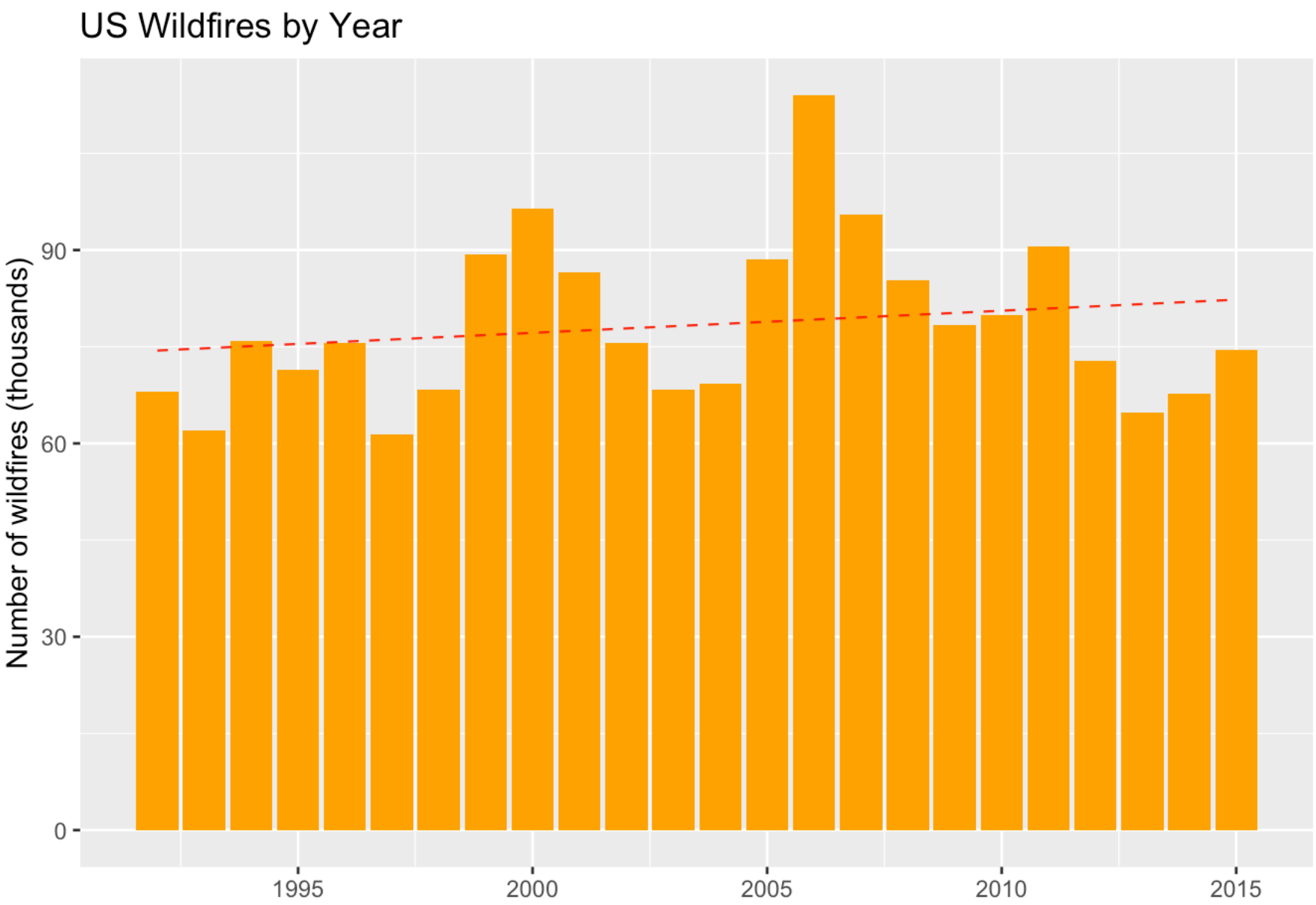
This database is pretty extensive. There is a lot of good stuff in here - spatial and temporal data. Let's see if we can find out anything interesting about wildfires in the US.

Wild fire over time

Annual | Day of the year | Daily | Monthly

undefined undefined undefined undefined

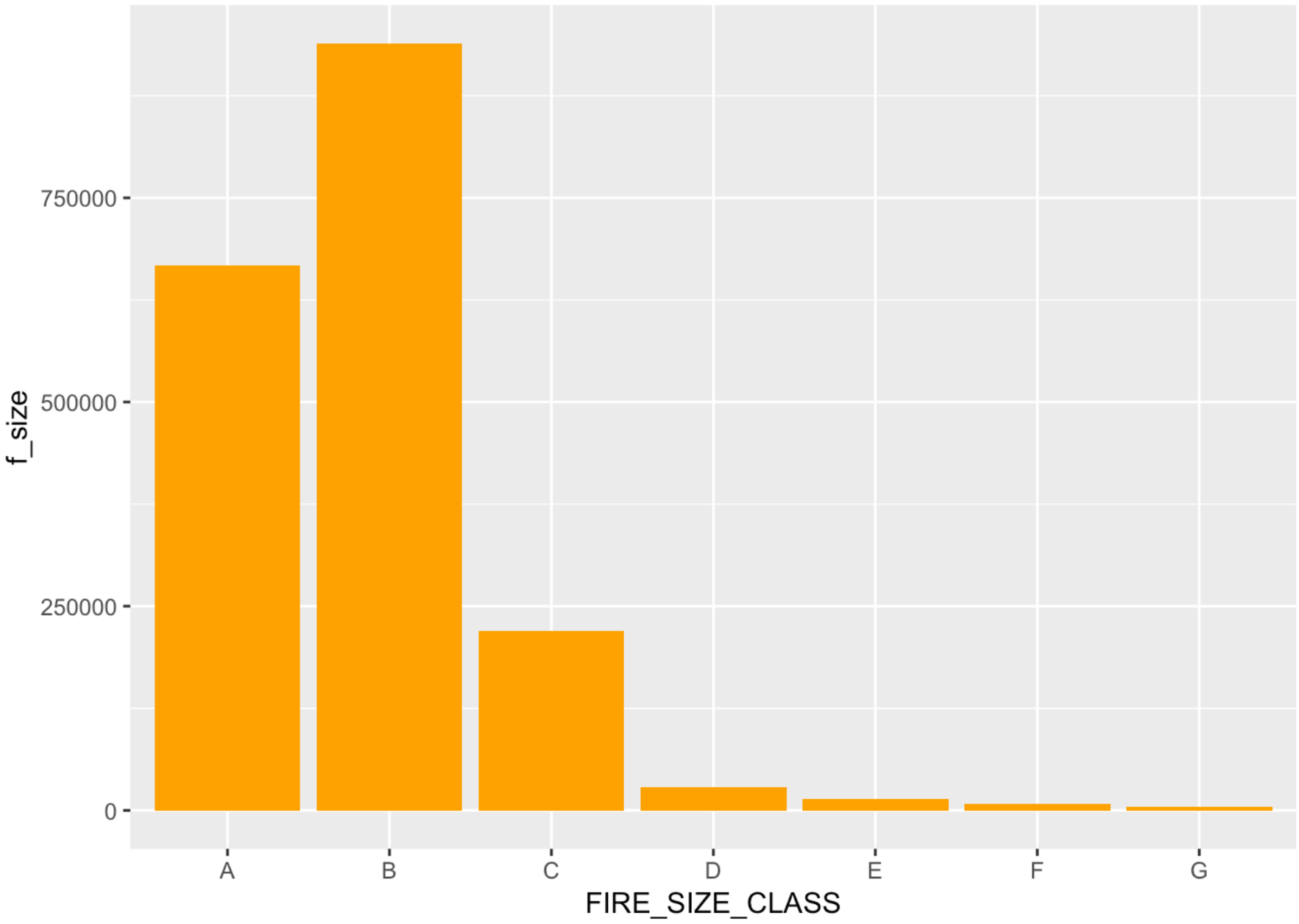
```
# fire ove the years

fires %>%
  group_by(FIRE_YEAR) %>%
  summarise(n_fires = n()) %>%
  ggplot(aes(x= FIRE_YEAR,y = n_fires/1000)) +
  geom_bar(stat = 'identity', fill = 'orange')+
  geom_smooth(method = 'lm', se = FALSE, linetype = 'dashed', size = 0.4, color = 're
d') +
  labs(x = '', y = 'Number of wildfires (thousands)', title = 'US Wildfires by Year')
```



The number of fires per year ran between 60,000 and 100,000 from 1992 to 2015. There was a spike in fires in 2006 to about 114,000. There is a small upward trend during this time period.

Fires by size

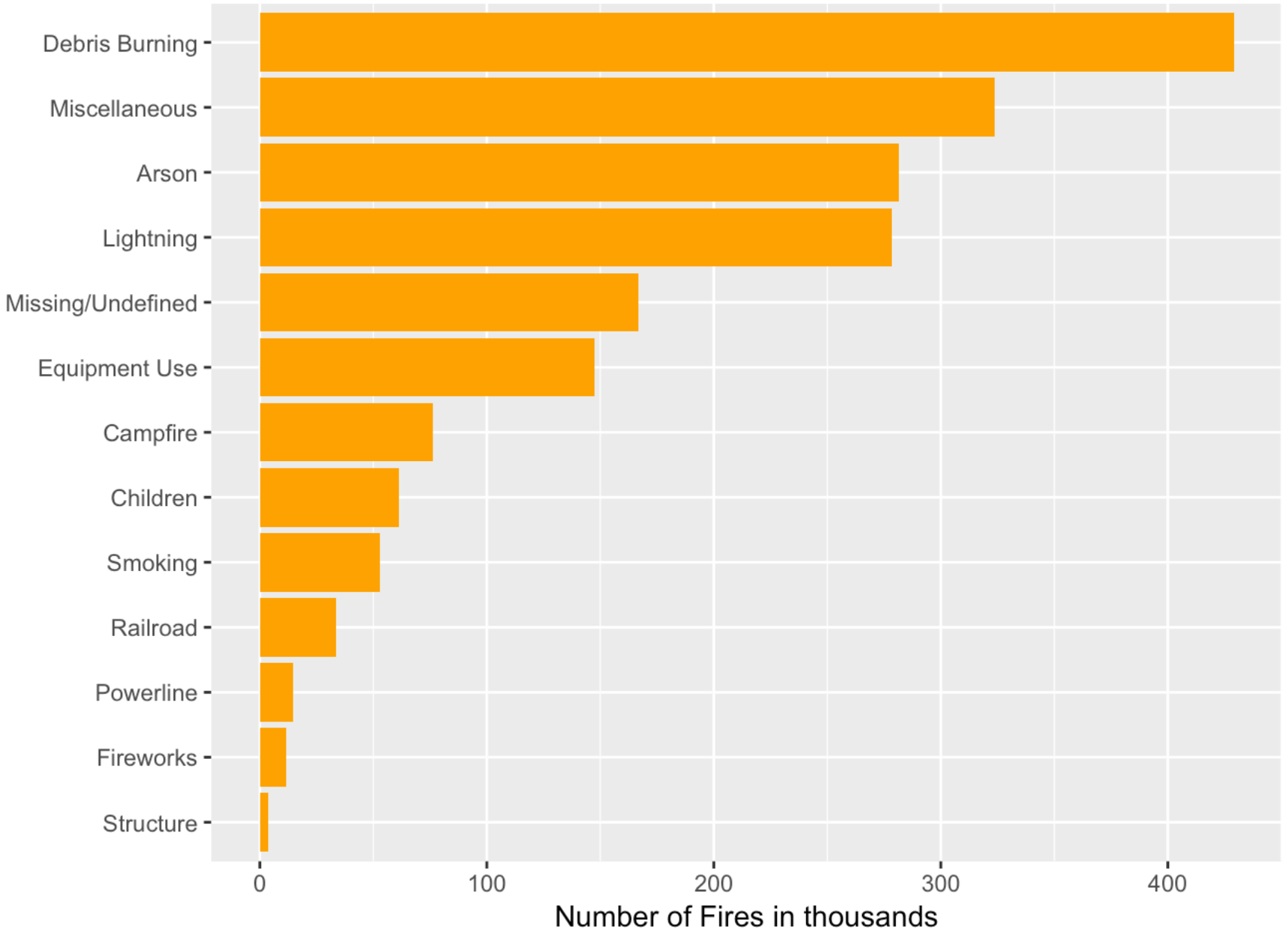
```
fires %>%
  group_by(FIRE_SIZE_CLASS) %>%
  summarise(f_size= n())%>%
  ggplot(aes(x = FIRE_SIZE_CLASS,y = f_size))+
  geom_bar(stat ='identity',fill = "orange")
```



Causes

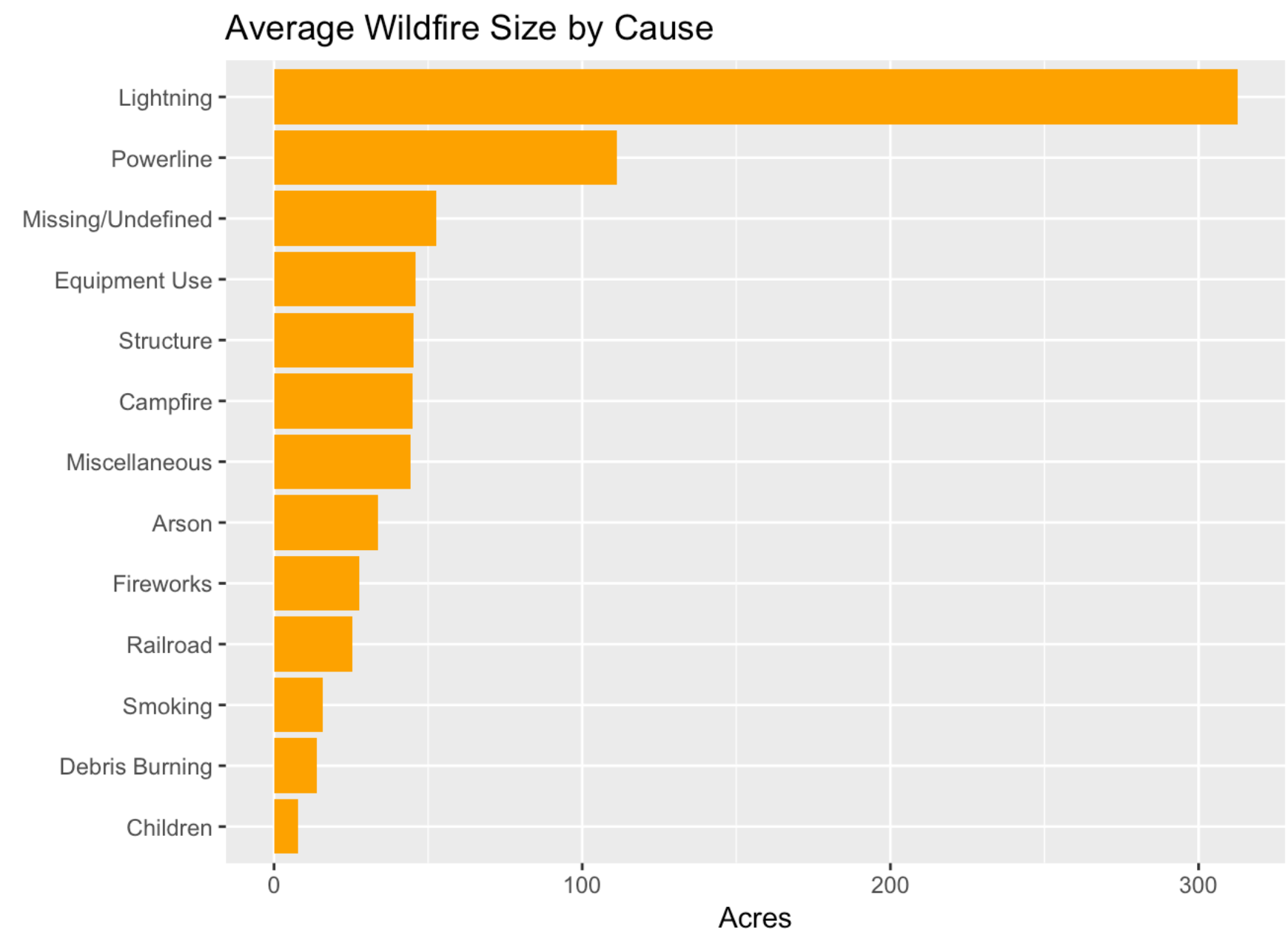
It would be interesting to examine the attributes of fires by cause. What causes the most fires? Which causes are associated with larger and longer-burning wildfires?

```
fires %>%
  group_by(STAT_CAUSE_DESCR) %>%
  summarise(n_reason = n()/1000) %>%
  ggplot(aes(x = reorder(STAT_CAUSE_DESCR,n_reason),y = n_reason ))+
  geom_bar(stat = "identity",fill= "orange")+
  coord_flip()+
  labs(x = "",y= "Number of Fires in thousands", tile = "Fire by cause")
```



Size of the fire by cause

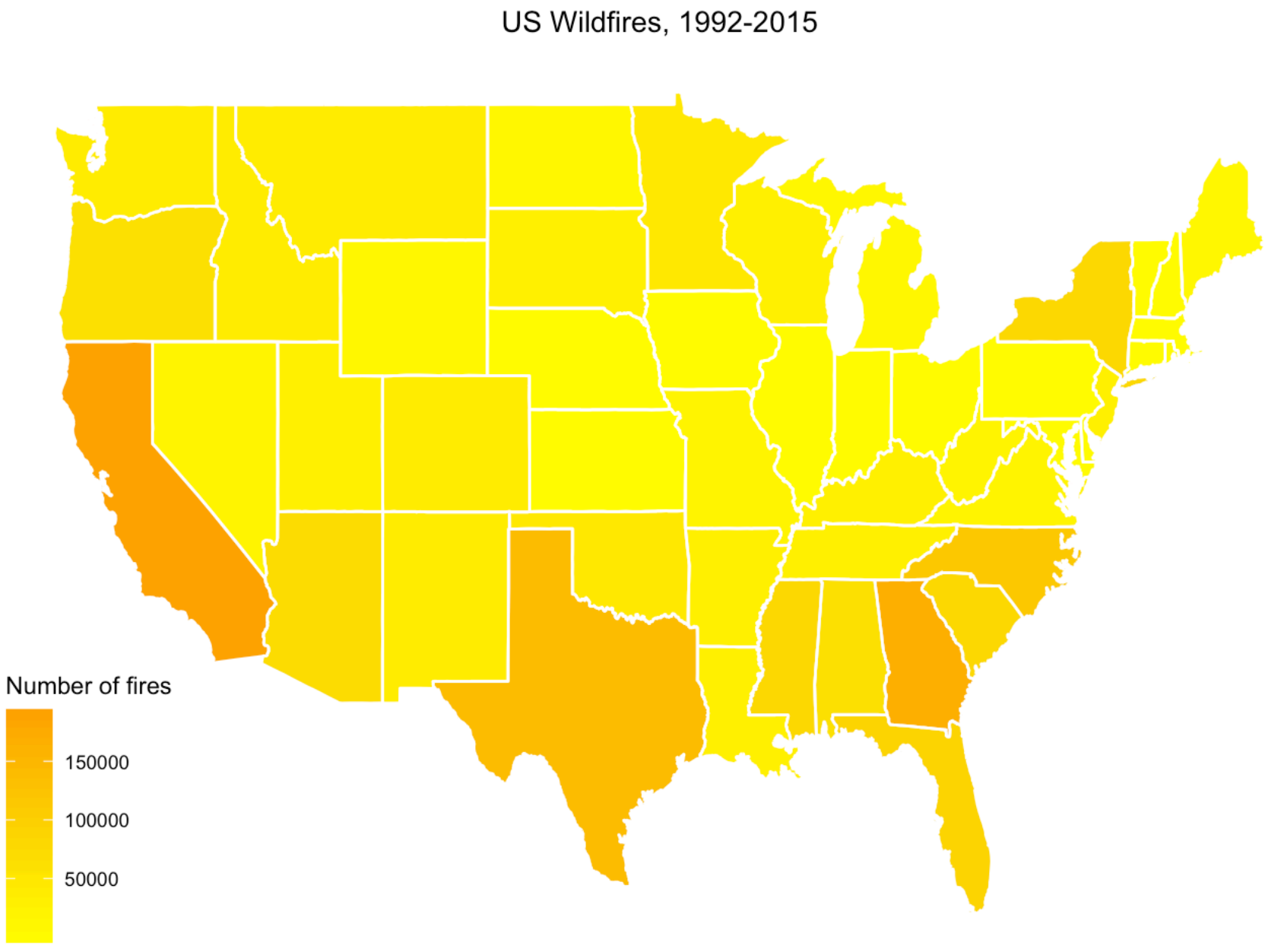
```
fires %>%
  group_by(STAT_CAUSE_DESCR) %>%
  summarise(mean_size = mean(FIRE_SIZE,na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(STAT_CAUSE_DESCR, mean_size), y = mean_size)) +
    geom_bar(stat = 'identity', fill = 'orange') +
    coord_flip() +
    labs(x = '', y = 'Acres', title = 'Average Wildfire Size by Cause')
```



Wildfire Geography

```
# Add codes for DC and Puerto Rico to the default state lists
state.abb <- append(state.abb, c("DC", "PR"))
state.name <- append(state.name, c("District of Columbia", "Puerto Rico"))

# Map the state abbreviations to state names so we can join with the map data
fires$region <- map_chr(fires$STATE, function(x) { tolower(state.name[grep(x, state.a
bb)]) })

# Get the us state map data
state_map <- map_data('state')
```

```
fires %>%
    select(region) %>%
    group_by(region) %>%
    summarize(n = n()) %>%
    right_join(state_map, by = 'region') %>%
    ggplot(aes(x = long, y = lat, group = group, fill = n)) +
    geom_polygon() +
    geom_path(color = 'white') +
    scale_fill_continuous(low = "yellow",
                          high = "orange",
                          name = 'Number of fires') +
    theme_map() +
   # ggplot2::coord_map('albers', lat0=30, lat1=40) +
    ggtitle("US Wildfires, 1992-2015") +
    theme(plot.title = element_text(hjust = 0.5))
```



surprised to see Georgia with so many fires. A map of wildfires normalized by size would be more interesting. I'll do that shortly. First let's look at fire causes by state.

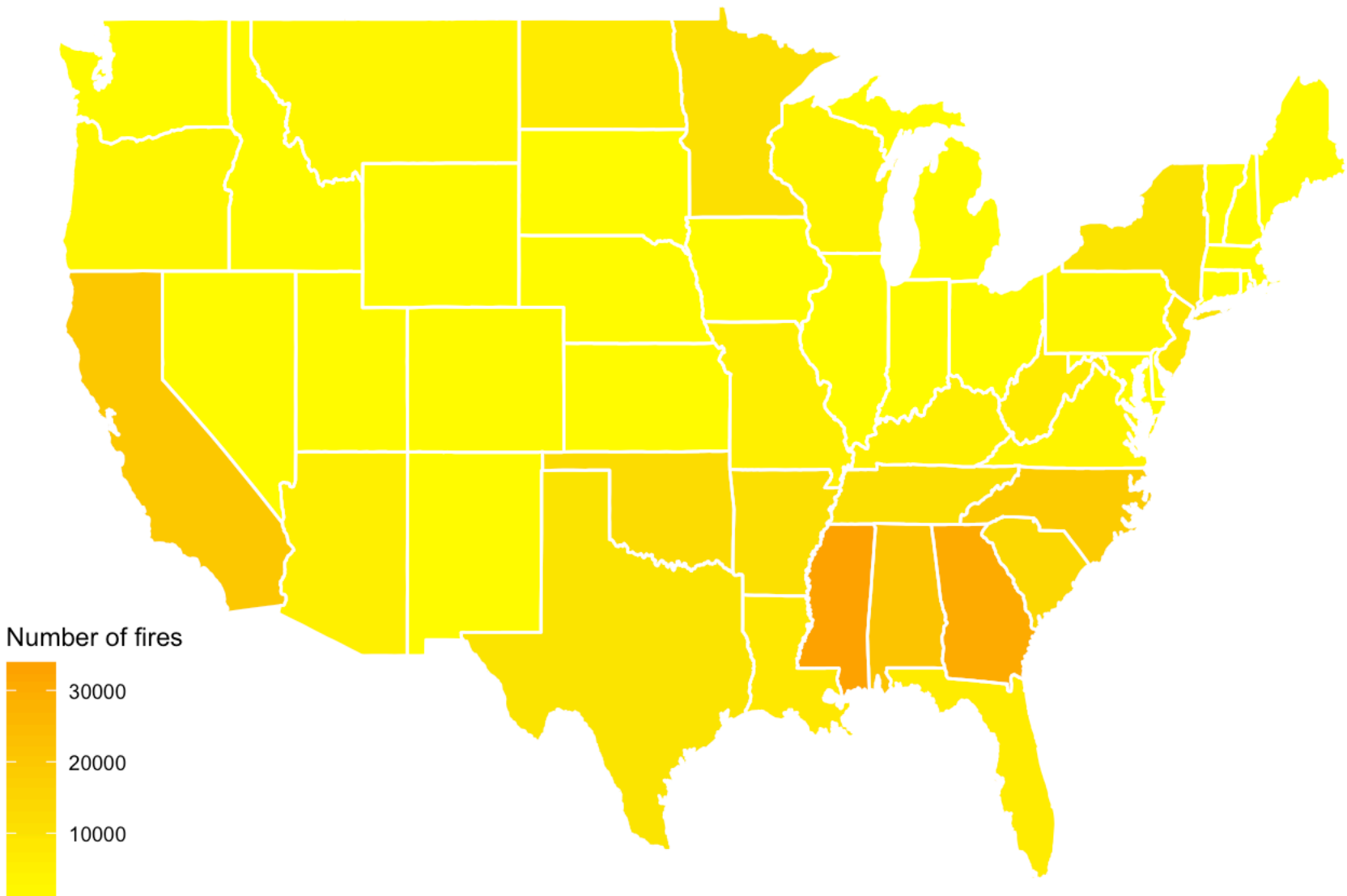
I'd like to make the same map for each of the fire causes. Because it will require using the same basic code block repeatedly, I will make it a function that we can reuse.

```
plotState <- function(cause){
  fires %>%
        filter(STAT_CAUSE_DESCR == cause) %>%
        select(region) %>%
        group_by(region) %>%
        summarize(n = n()) %>%
        right_join(state_map, by = 'region') %>%
        ggplot(aes(x = long, y = lat, group = group, fill = n)) +
        geom_polygon() +
        geom_path(color = 'white') +
        scale_fill_continuous(low = "yellow",
                              high = "orange",
                              name = 'Number of fires') +
        theme_map() +
        ggtitle(paste0("US Wildfires Caused by ", cause, ", 1992-2015")) +
        theme(plot.title = element_text(hjust = 0.5))

}
```

Fires by state

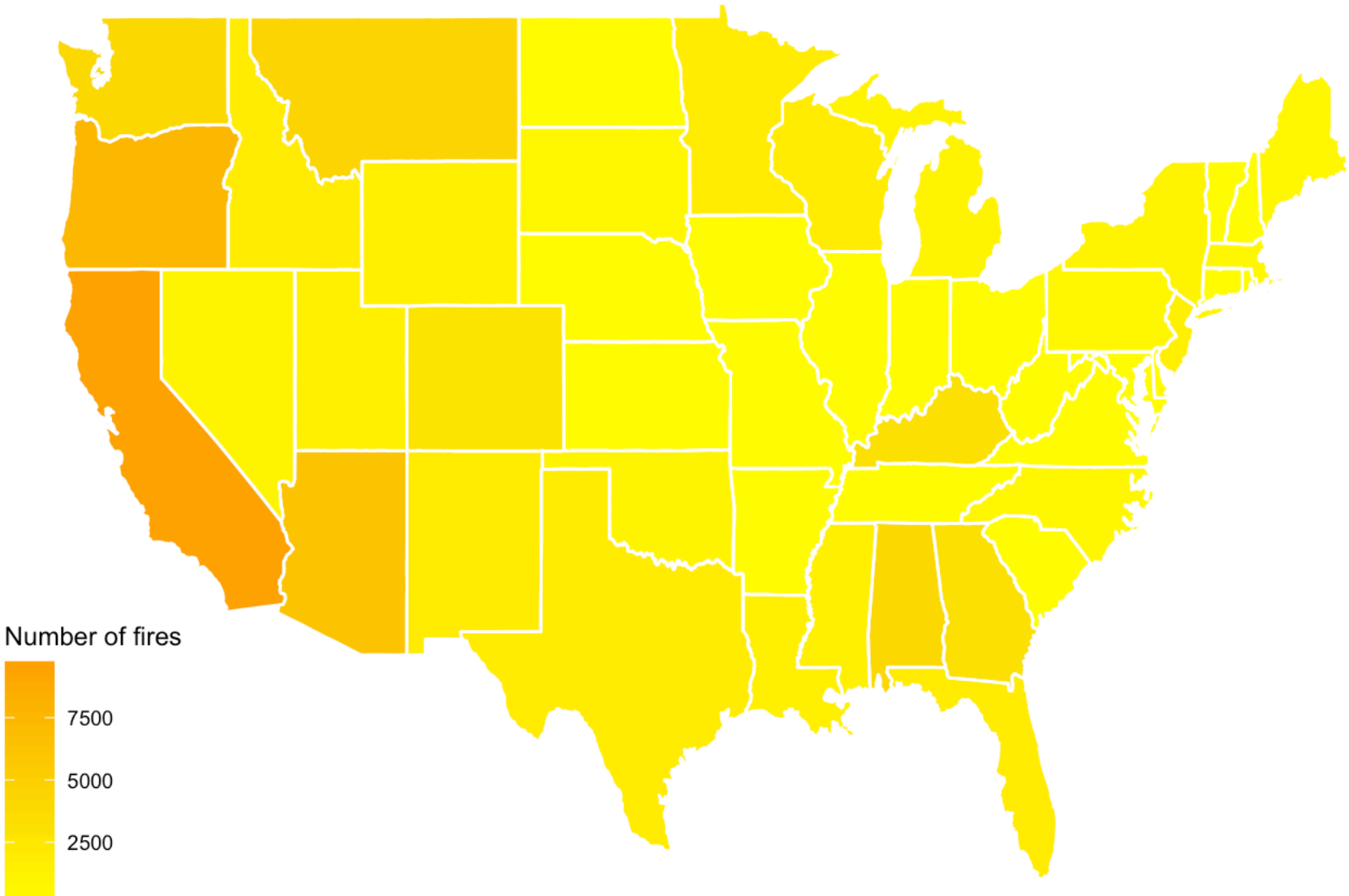
Total	Fireworks	Lightning
undefined	undefined	undefined

```
plotState(cause = "Arson")
```



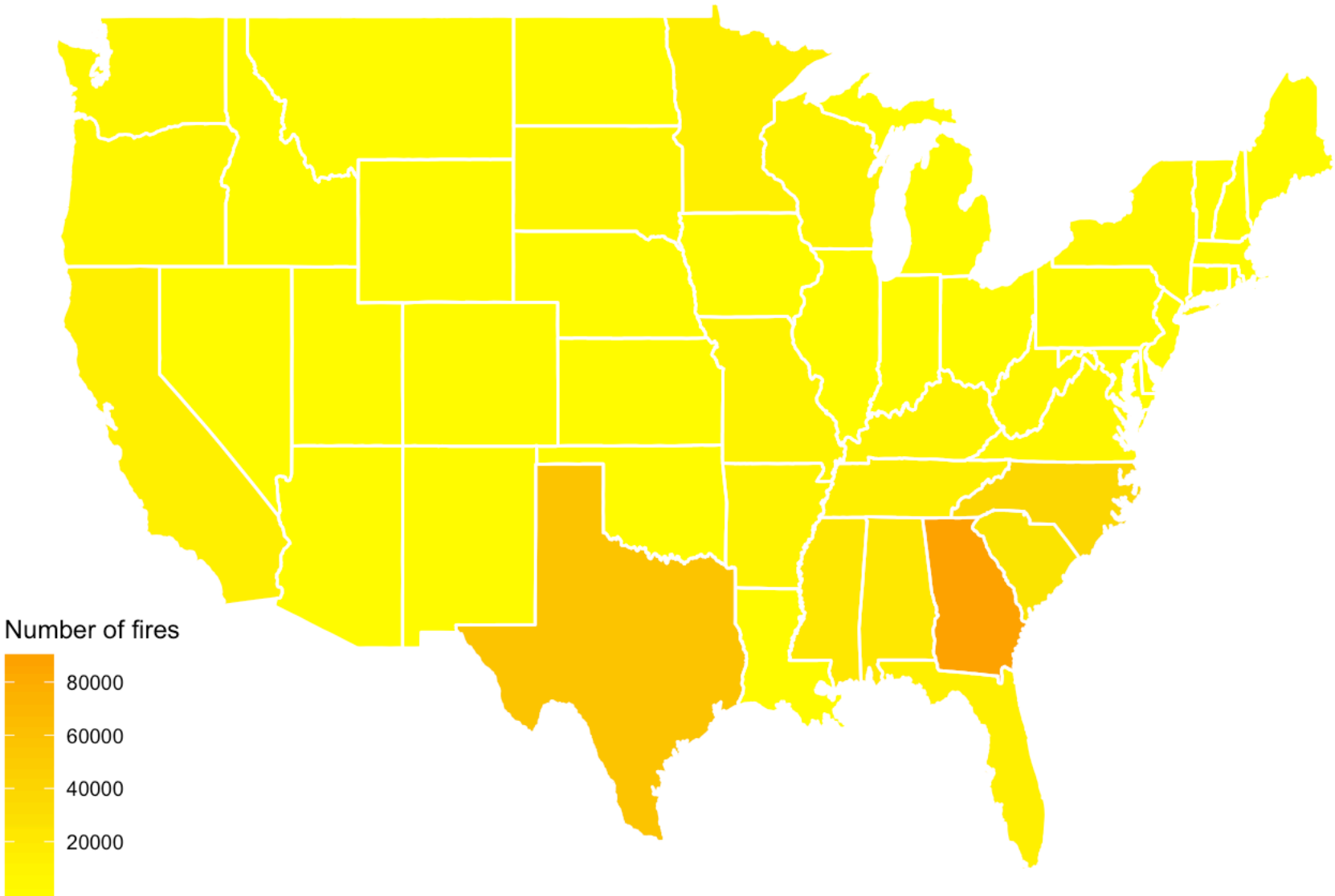
Campfire

```
plotState(cause = "Campfire")
```



Debris Burning

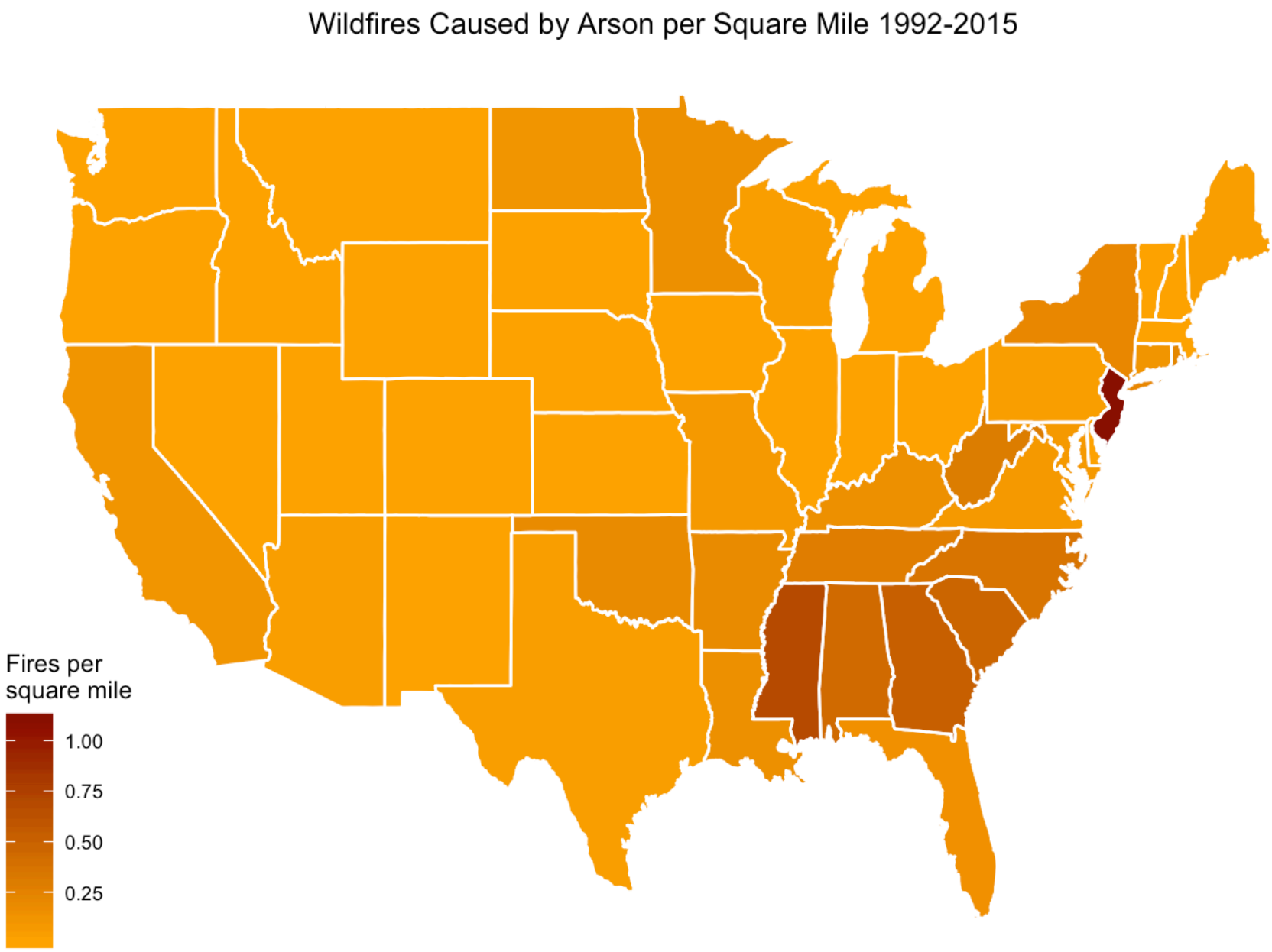
```
plotState(cause = "Debris Burning")
```



Arson | Campfire | Lightning | Debris Burning

undefined | undefined | undefined | undefined

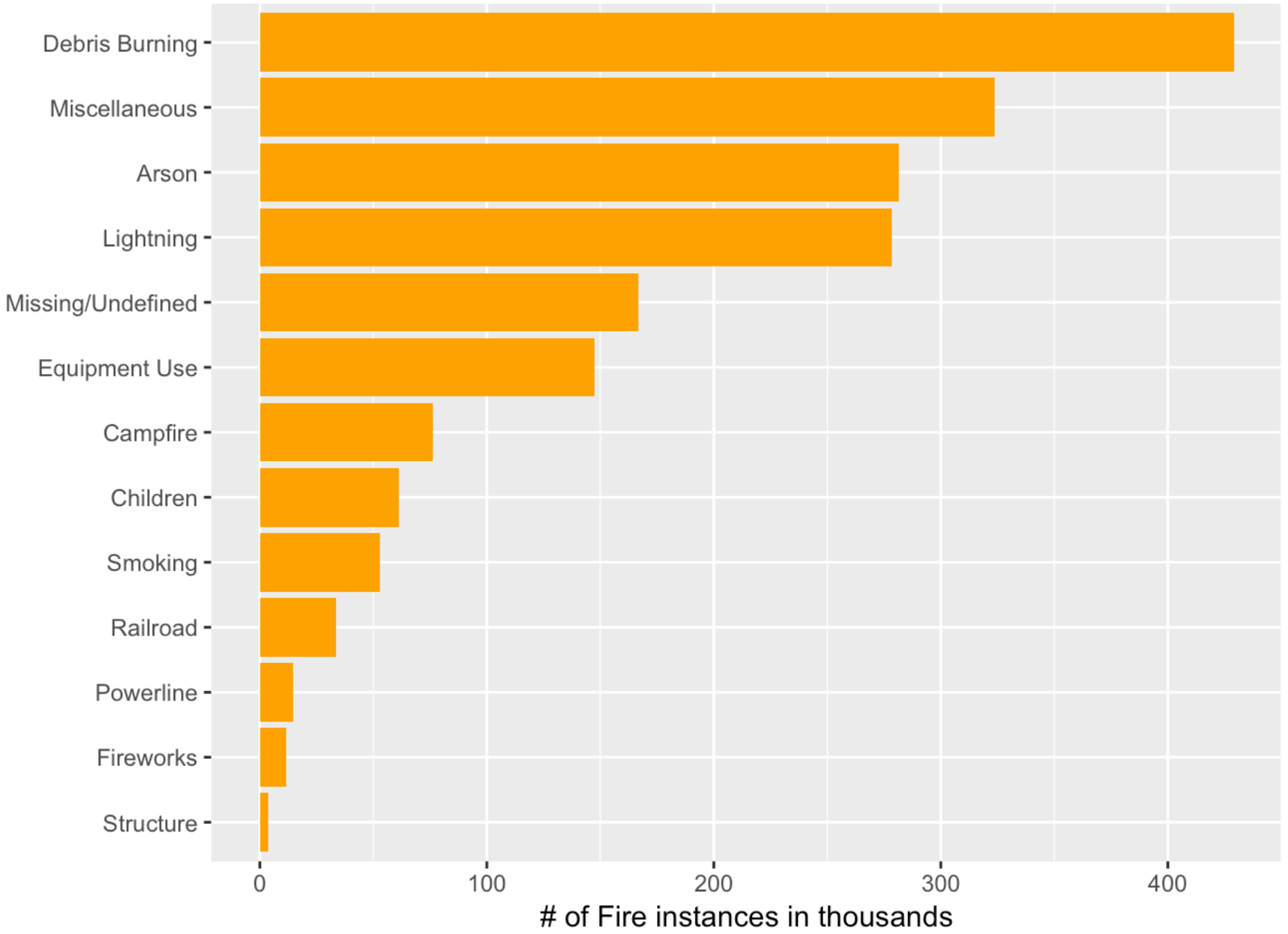
```
plotNormalizedState(cause = "Arson")
```



Target Feature Analysis

First, let's take a look at what we are trying to predict. The column `STAT_CAUSE_DESCR` has the fire cause. We want to know what these are and how they are distributed.

```
fires %>%
  group_by(STAT_CAUSE_DESCR)%>%
  summarise(n_dist = n()) %>%
  ggplot(aes(x=reorder(STAT_CAUSE_DESCR,n_dist),y=n_dist/1000))+
  geom_bar(stat='identity',fill = "orange")+
  coord_flip()+
  labs(x="",y= "# of Fire instances in thousands")
```



‘Debris Burning’ is the most common cause by far in this sample. ‘Miscellaneous’, ‘Lightning’, and ‘Arson’ are fairly prevalent as well. At the other end we see some causes that are far less common. Because their frequency is so low, we may run into difficulty in predicting these classes.

Data Setup

First, let’s choose what features we want to use in a model. Then we’ll split our data into a train and test set. To start, let’s choose only a single feature , `FIRE_SIZE` for simplicity’s sake.

```
# features to use
features <- c('FIRE_SIZE')

fires$STAT_CAUSE_DESCR <- as.factor(fires$STAT_CAUSE_DESCR)

# index for train/test split
set.seed(123)
train_index <- sample(c(TRUE, FALSE), nrow(fires), replace = TRUE, prob = c(0.8, 0.2)
)
test_index <- !train_index

# Create x/y, train/test data
x_train <- as.data.frame(fires[train_index, features])
y_train <- fires$STAT_CAUSE_DESCR[train_index]

x_test <- as.data.frame(fires[test_index, features])
y_test <- fires$STAT_CAUSE_DESCR[test_index]
```

Iteration 1: Benchmark

Before we start modelling we should set a benchmark for ourselves. If our model is not more accurate than a benchmark, then our fancy modeling is all for naught. In this case, a simple benchmark might be to just always predict the most common class - 'Debris Burning'. Let's see how accurate this method is on our test data. Note that this is equivalent to calculating the percent of the test data labeled 'Debris Burning'.

```
preds <- rep('Debris Burning', length(y_test))

test_set_acc <- round(sum(y_test == preds)/length(preds), 4)
print(paste(c("Accuracy:" , test_set_acc)))
```

```
## [1] "Accuracy:" "0.2276"
```

This naive model has an accuracy of about 22.9%. Surely we can do better than that.

Iteration 2: A Simple Decision Tree

We'll start with a simple decision tree. Rather than use the `rpart` package directly, we'll use it through `caret`. Whenever possible, I highly recommend using `caret` for most ML tasks in R since it provides a common API for using many different model types that are scattered throughout R and its numerous packages. Let's train this decision tree using our lonely `FIRE_SIZE` feature.

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# create the training control object.
tr_control <- trainControl(method = 'cv', number = 3)

# Train the decision tree model
set.seed(123)
dtree <- train(x = x_train,
               y = y_train,
               method = 'rpart',
               trControl = tr_control)
```

```
pred <- predict(dtree,newdata = x_test)

#calculate the model accuracy

test_set_acc <- round(sum(y_test==pred)/length(pred),4)
print(paste(c("Accuracy :",test_set_acc)))
```

```
## [1] "Accuracy :" "0.2699"
```

The accuracy of our simple decision tree model yields 27.1% accuracy on our test set. It appears we've already beat our benchmark but we should be careful as we don't really know by how much this score will vary on other random test sets. To get further intuition, we can examine the scores on the holdout sets used during cross-validation:

```
print(dtree$resample)
```

```
##    Accuracy      Kappa Resample
## 1 0.2748420 0.08607568    Fold3
## 2 0.2747557 0.08603340    Fold2
## 3 0.2696408 0.08245400    Fold1
```

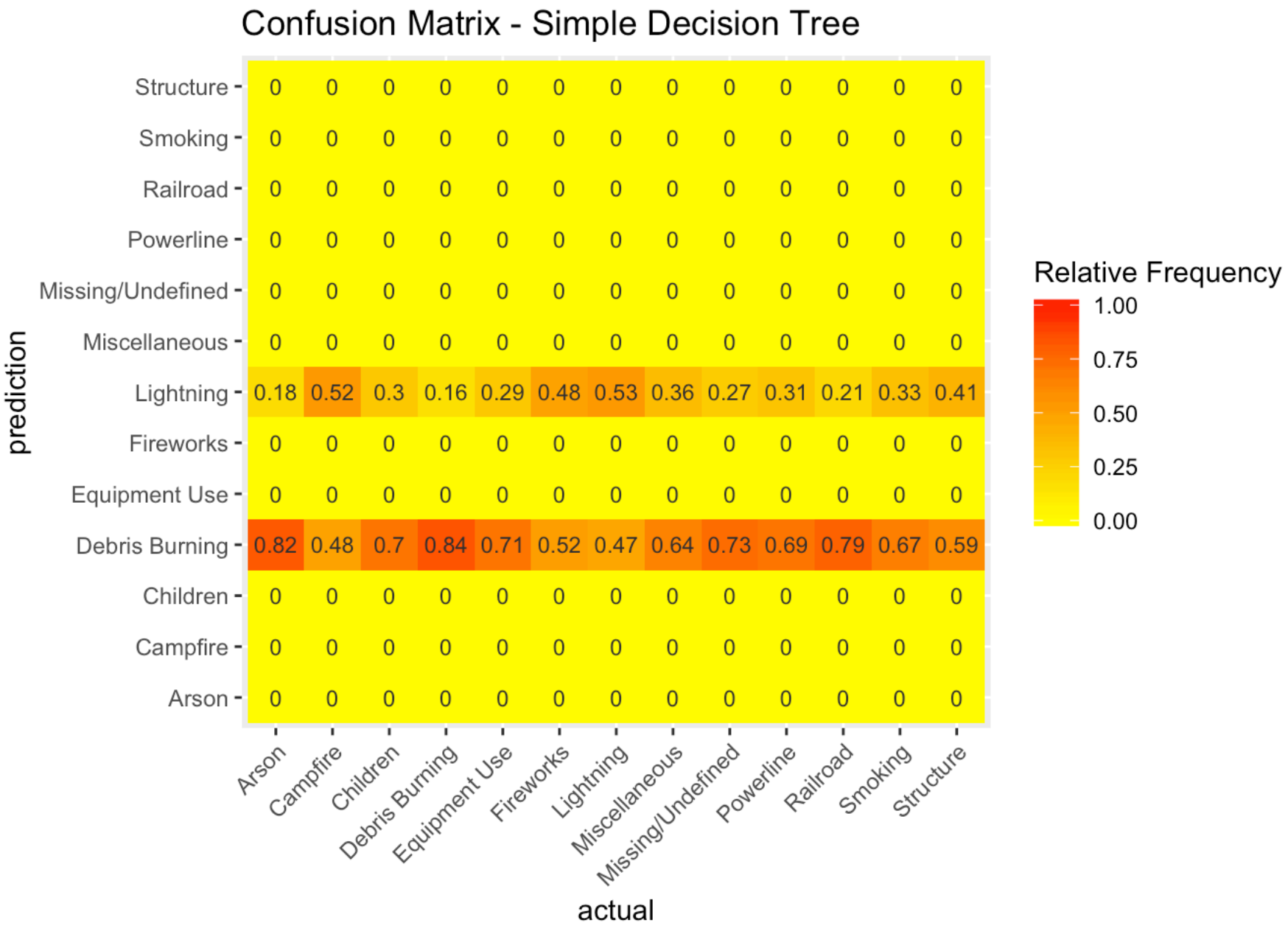
It looks like the accuracy score was similar during cross-validation. A good sign, but we should look deeper still. Accuracy is a fairly simple metric that will often not be able to capture the nuances of multi-class classification. Let's take a look at the confusion matrix. Because we have thirteen possible classes, this confusion matrix will be rather large, so let's dress it up a bit:

Results

Confusion Matrix Heatmap | Tree Plot | CV Plot

undefined | undefined | undefined

```
library(tibble)
confusionMatrix(y_test,pred)$table %>%
  prop.table(margin = 1) %>%
  as.data.frame.matrix() %>%
  rownames_to_column(var="actual") %>%
  tidyr::gather(key = "prediction",value = "freq",-actual) %>%
   ggplot(aes(x = actual, y = prediction, fill = freq)) +
    geom_tile() +
    geom_text(aes(label = round(freq, 2)), size = 3, color = 'gray20') +
    scale_fill_gradient(high = 'Red', low = 'Yellow', limits = c(0,1), name = 'Relati
ve Frequency') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    ggtitle('Confusion Matrix - Simple Decision Tree')
```



Notice from the confusion matrix plot, table, and the tree plot that our model is currently predicting only two of the thirteen classes.

Iteration 3 : More Features

Let's include more features in the training data. Here we add the `FIRE_YEAR` and the `DISCOVERY_DOY` feature, which is the day of the year that the fire was discovered.

```
features <- c("FIRE_YEAR","DISCOVERY_DOY","FIRE_SIZE")
x_train <- as.data.frame(fires[train_index,features])
y_train <- fires$STAT_CAUSE_DESCR[train_index]

x_test <- as.data.frame(fires[test_index,features])
y_test <-fires$STAT_CAUSE_DESCR[test_index]
```

```
# Train tree model 2

set.seed(123)

dtree <-train(x=x_train,
              y=y_train,
              method = "rpart",
              tuneLength = 5,
              trControl = tr_control)
```

```
preds <- predict(dtree,newdata = x_test)

# Accuracy of the test data

Accuracy <- sum(y_test==preds)/length(preds)
print(paste(c("Accuracy:",round(Accuracy,4))))
```

```
## [1] "Accuracy:" "0.3228"
```

The accuracy score on the test set has improved. Again let's take a look at the cross-validation scores to see if the results are similar:

```
print(dtree$resample)
```

```
##    Accuracy     Kappa Resample
## 1 0.3237376 0.1636490    Fold1
## 2 0.3243518 0.1642466    Fold2
## 3 0.3225639 0.1648027    Fold3
```

Great. They are consistent with our test set score.

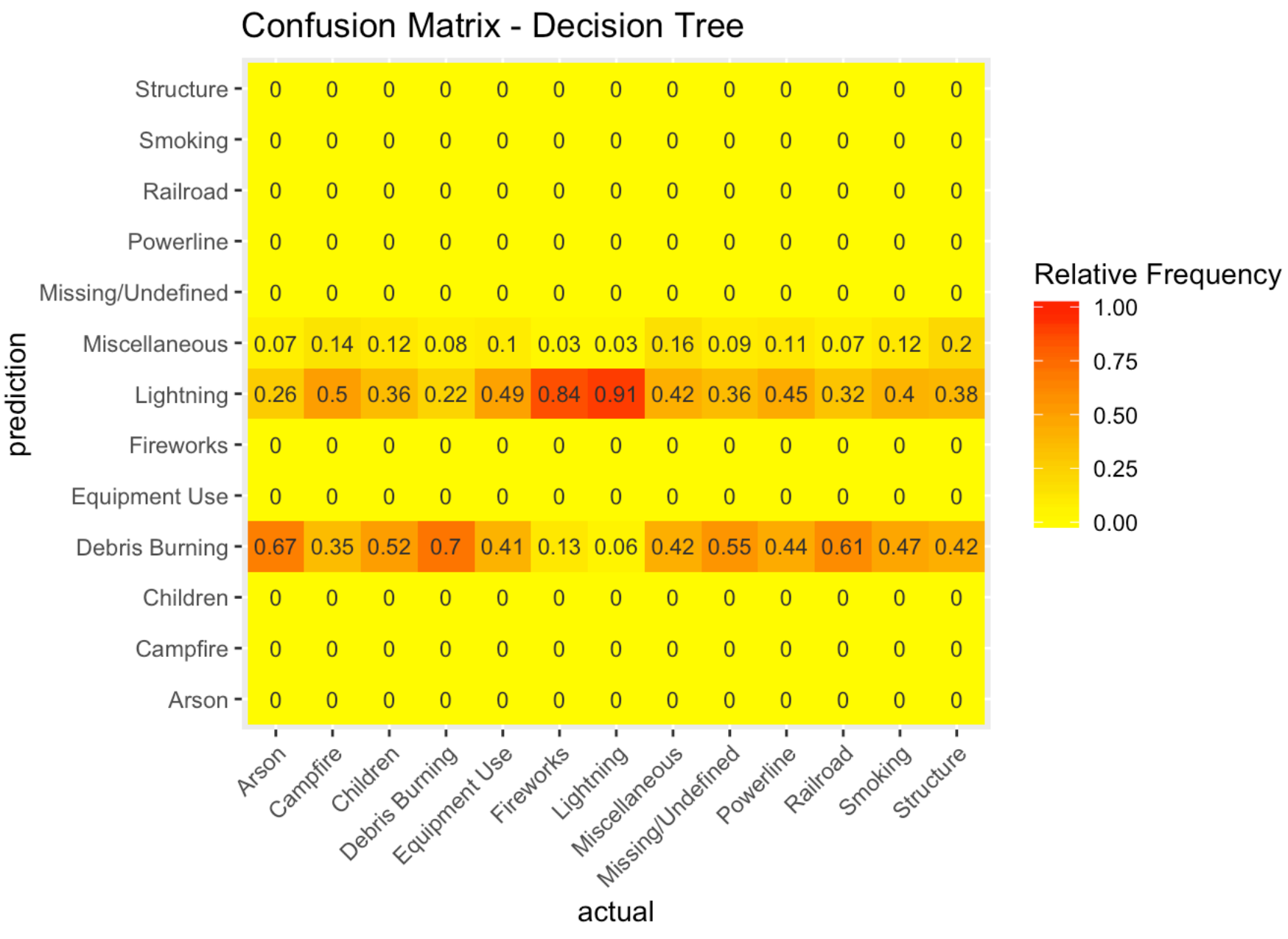
Let's take a look at the new confusion matrix:

Results

Confusion Matrix Heatmap | Tree Plot | CV plot | Results

undefined | undefined | undefined | undefined

```
confusionMatrix(y_test,preds)$table %>%
  prop.table(margin = 1) %>%
  as.data.frame.matrix() %>%
  rownames_to_column(var = 'actual') %>%
  tidyr::gather(key = "prediction",value ="freq",-actual) %>%
  ggplot(aes(x=actual,y=prediction,fill= freq))+
  geom_tile()+
  geom_text(aes(label = round(freq, 2)), size = 3, color = 'gray20') +
    scale_fill_gradient(low = 'yellow', high = 'red', limits = c(0,1), name = 'Relati
ve Frequency') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    ggtitle('Confusion Matrix - Decision Tree')
```




```
# show confusion matrix
confusionMatrix(y_test, preds)$table %>%
    as.data.frame.matrix()
```

```
##                   Arson Campfire Children Debris Burning Equipment Use
## Arson                 0        0        0           37429             0
## Campfire              0        0        0            5382             0
## Children              0        0        0            6375             0
## Debris Burning        0        0        0           59962             0
## Equipment Use         0        0        0           11927             0
## Fireworks             0        0        0             285             0
## Lightning             0        0        0            3248             0
## Miscellaneous         0        0        0           27488             0
## Missing/Undefined     0        0        0           18298             0
## Powerline             0        0        0            1256             0
## Railroad              0        0        0            4074             0
## Smoking               0        0        0            4940             0
## Structure             0        0        0             314             0
##                   Fireworks Lightning Miscellaneous Missing/Undefined
## Arson                     0     14666          4175                 0
## Campfire                  0      7712          2205                 0
## Children                  0      4442          1424                 0
## Debris Burning            0     19066          6545                 0
## Equipment Use             0     14550          2969                 0
## Fireworks                 0      1885            64                 0
## Lightning                 0     51129          1618                 0
## Miscellaneous             0     27179         10270                 0
## Missing/Undefined         0     11888          2989                 0
## Powerline                 0      1296           320                 0
## Railroad                  0      2174           479                 0
## Smoking                   0      4216          1266                 0
## Structure                 0       284           147                 0
##                   Powerline Railroad Smoking Structure
## Arson                     0        0       0         0
## Campfire                  0        0       0         0
## Children                  0        0       0         0
## Debris Burning            0        0       0         0
## Equipment Use             0        0       0         0
## Fireworks                 0        0       0         0
## Lightning                 0        0       0         0
## Miscellaneous             0        0       0         0
## Missing/Undefined         0        0       0         0
## Powerline                 0        0       0         0
## Railroad                  0        0       0         0
## Smoking                   0        0       0         0
## Structure                 0        0       0         0
```

```
#kable("html") %>%
#kable_styling(bootstrap_options = c('striped'), font_size = 8) %>%
#scroll_box(height = "400px")
```

Interesting. Now our model is predicting ‘Miscellaneous’ in addition to ‘Debris Burning’ and ‘Lightning’. For every $27488 + 27179 + 1618 = 56285$ times that the fire’s cause was ‘Miscellaneous’, the model got it right 10270 times. This isn’t a very interesting class though. I’d really like to see a model that can predict Arson with a reasonable level of accuracy.