An Analysis of Wildfire Size Based on Size Class

Wyatt Workman (Waworkman@ucdavis.edu (mailto:Waworkman@ucdavis.edu)), Selam Berekat(smberekat@ucdavis.edu (mailto:smberekat@ucdavis.edu))

6/4/2021

Contributions:

Selam:

Analyzed which factors lead to size F and G wildfires by creating plots, wrote conclusion.

Wyatt:

Initial subsetting of dataset, form initial logistic model for predicting class F or G, compare this method to KNN and LDA.

Introduction:

As climate change progresses, natural disasters will become more frequent and more intense. In California, and much of the western United States, this means more frequent and intense wildfires, as we witnessed during the summer of 2020. Our firefighters and forest managers will be much more successful in extinguishing these fires if they have a way of predicting the size and occurrence of a wildfire based on current weather conditions.

Therefore, the main objective of this project is to develop a model that predicts the size class of a wildfire based on several factors, including temperature, wind, humidity, precipitation, location, and cause, before a fire occurs. In addition, we will also explore the extent to which type of vegetation affects wildfire size, as well as how much the size of a wildfire affects the amount of time it takes to extinguish it.

Dataset:

Our dataset was found on Kaggle.com. It contains approximately 56,000 rows and 43 different columns. The dataset includes information on fires from 1992-2015, including size, putout time, weather parameters prior to the start of the fire, location, and cause of the fire. We eliminated approximately half of the rows, as there were many rows that contained incorrect values that did not make logical sense for our analysis (example: you cannot have negative humidity or wind speed).

Research questions:

1. Construct a model to predict whether a fire will be a class F or G wildfire?

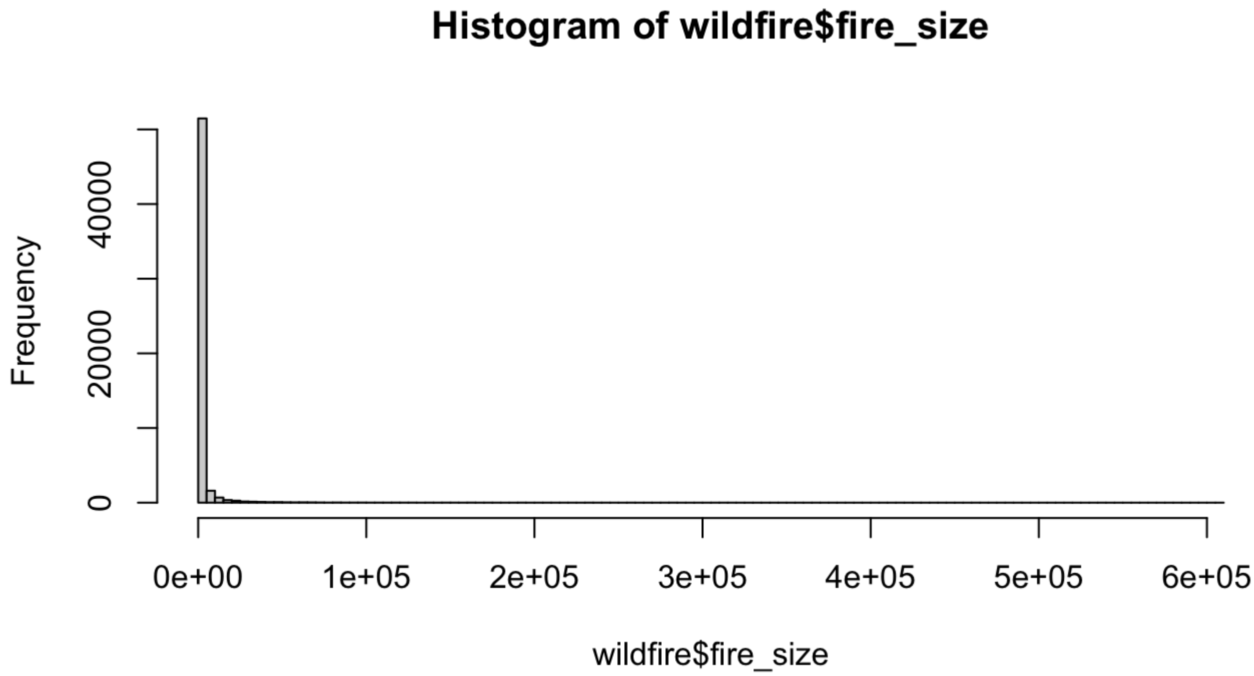
2. Which factors are important in predicting a class F or G wildfire?

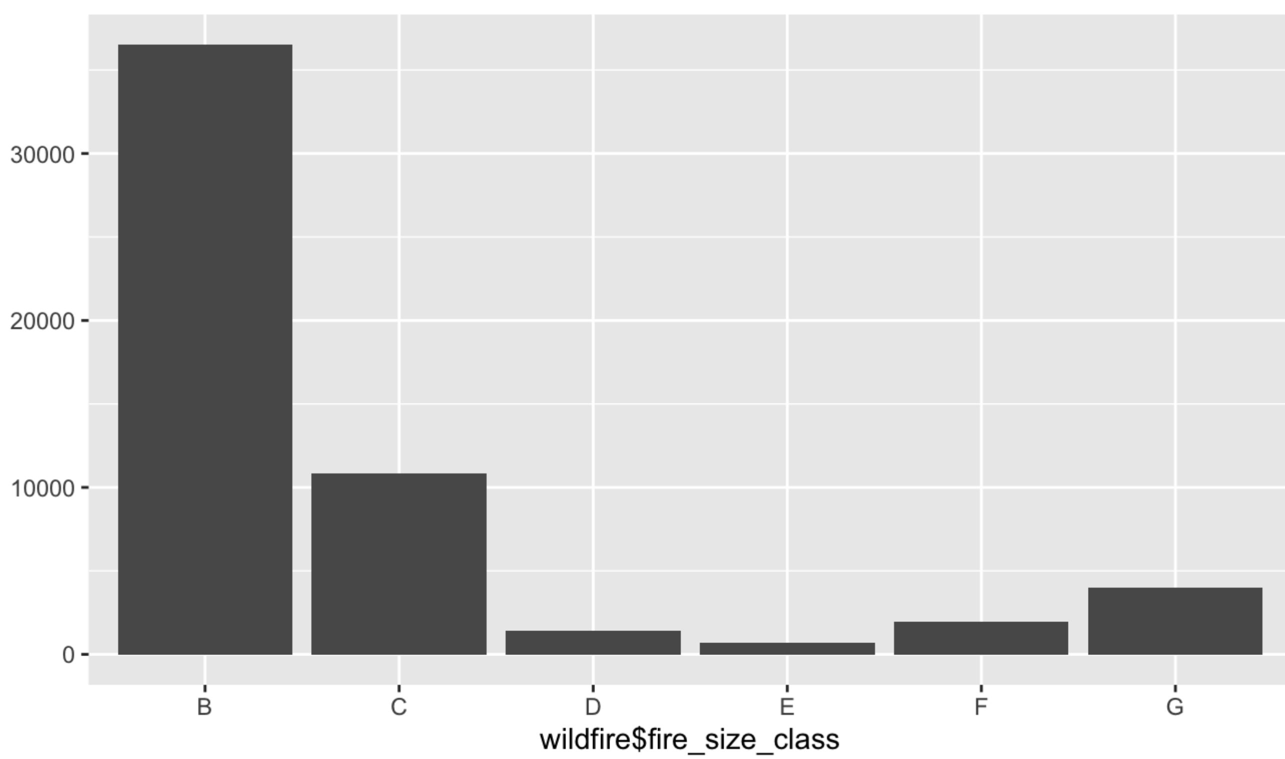
3. What factors cause type F and G wildfires?

Methods:

In order to formulate a model that will classify fire size class, we will employ a logistic regression to determine optimal model predictors and to classify the fire size class, as well as an LDA and KNN procedure to determine which yields the lowest error rate, and thus most accurately classifies a wildfire as class F or G.

Explore the distribution of two possible response variables:





The first histogram represents the distribution of our data is skewed to the left and indicates there are values that have occurred most often, and this is because we have discrete variables. The second bar plot represents the response variable fire size class. It suggests that the level of the fire size class B having an increased observation following with C and G. And the observation for fire size class E is the lowest level with observation. In this analysis, we will analyze fire classes F and G, as these are the largest fire types and would require the most resources to extinguish.

After performing a stepwise logistic regression, the optimal model is:

```
## as.numeric(fire_size_class) ~ Temp_pre_30 + Temp_pre_7 + Prec_pre_30 +
##     Prec_pre_15 + remoteness + latitude + longitude + disc_pre_year +
##     stat_cause_descr
```

Predict the probability of a fire being of class F using the logistic model developed above (confusion matrix and error rate):

```
##          predicted.class
## true.resp    F    G
##         F   84  809
##         G   73 1719
```

```
## [1] 0.3284916
```

This model is 67% accurate at predicting the probability that a fire is of class F or G.

Determine which model parameters are not significant:

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: as.numeric(fire_size_class)
##
## Terms added sequentially (first to last)
##
##
##                             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                         2684     3415.3
## Temp_pre_30                  1   1.9842      2683     3413.3  0.158946
## Temp_pre_7                   1   3.9251      2682     3409.4  0.047570 *
## Wind_pre_30                  1   5.3839      2681     3404.0  0.020324 *
## Prec_pre_30                  1   7.1381      2680     3396.9  0.007546 **
## Prec_pre_15                  1   3.9957      2679     3392.9  0.045617 *
## remoteness                   1  24.0429      2678     3368.8 9.421e-07 ***
## latitude                     1   3.4510      2677     3365.4  0.063215 .
## longitude                    1   9.4941      2676     3355.9  0.002061 **
## disc_pre_year                1   5.5503      2675     3350.3  0.018478 *
## as.factor(stat_cause_descr) 12  30.8732      2663     3319.5  0.002060 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears that temp_pre_30 is not significant. Remove it and rerun the model to see if prediction accuracy improves:

```
## [1] 0.6711359
```

This model is just as accurate after removing some of the less significant predictors.

Now, compare the accuracy of the logistic regression classifier to that of KNN and LDA:

```
##    FG.knn
##       F   G
##   F 131  92
##   G 282 166
```

```
## [1] 0.557377
```

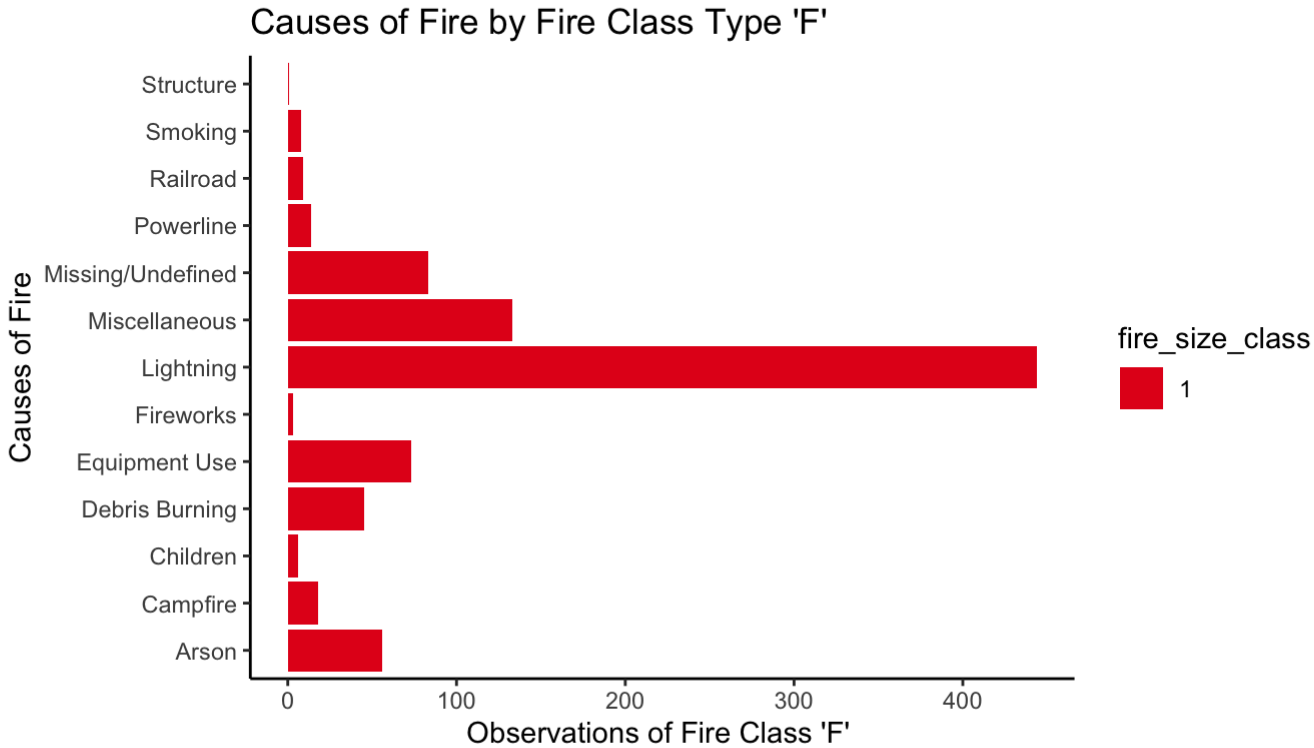
```
##
##       F   G
##   F 101 122
##   G 216 232
```

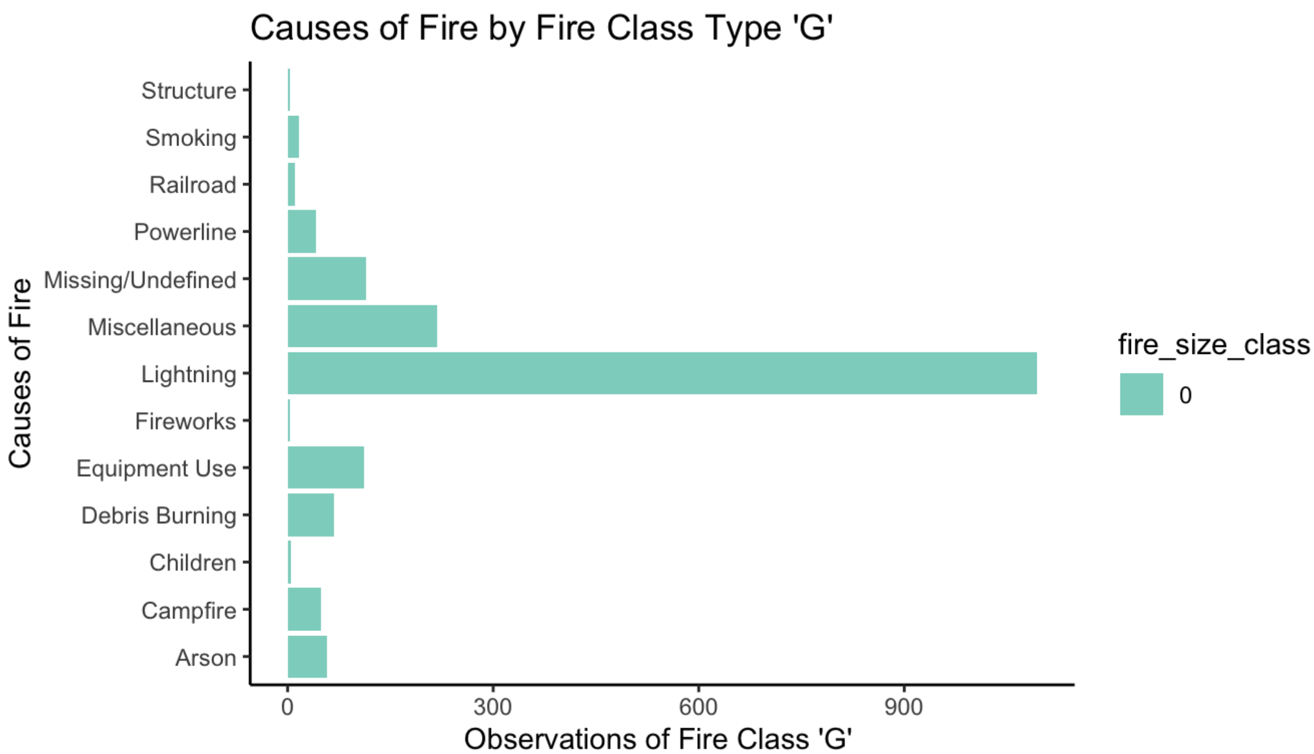
```
## [1] 0.5037258
```

From running the model using a logistic regression, KNN and LDA, we can conclude that using a logistic regression on classes F and G is the most accurate classifier of wildfire types F and G.

We now turn our attention to analyzing which factors influence the liklihood of a wildfire being of class size F or G:

```
##
##           Arson Campfire Children Debris Burning Equipment Use Fireworks
##   Fire_G_ 0.022    0.018    0.002          0.025         0.042     0.001
##   Fire_F_ 0.021    0.007    0.002          0.017         0.027     0.001
##
##           Lightning Miscellaneous Missing/Undefined Powerline Railroad Smoking
##   Fire_G_     0.407         0.081             0.043     0.015    0.004   0.006
##   Fire_F_     0.165         0.050             0.031     0.005    0.003   0.003
##
##           Structure
##   Fire_G_     0.001
##   Fire_F_     0.000
```





We conclude that most likely cause of wildfire is due to Lightning in both case of fire class type of F and type of G.

Conclusion:

From these two bar plots, we can see the relationship of a fire class either F or G and causes of those fires. Plot 1 of Type F fire illustrates Lightning have the highest risk of causing fire class F with 16.5%. in contrast, we can observe that smoking, fireworks or campfire have relatively smaller portion of causing fire F. Plot 2 of Type G fire also illustrates Lightning having the highest risk of causing fire class G with 40.7% chance to occur. While the other causes such as campfire, fireworks, or smoking have relatively small proportion to cause a fire of type G. Overall, we would consider the shape of our Lightning to have a higher probability of causing fire type of F and G.

Our objective was to develop a model that is able to classify a wildfire as a size class F or G wildfire. Using our dataset and analysis techniques learned in class, we were able to develop a logistic model that is 67% accurate at making these predictions. We also compared this model to two other classification techniques: k-nearest neighbors and linear discriminant analysis, in order to determine if one of these classifiction techniques would yield a more accurate prediction rate. We conclude that, in the case of our data, the logistic regression model is the most accurate way to classify type F and G wildfires.

Code Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
wildfire = data.frame(read.csv("FW_Veg_Rem_Combined.csv"))
attach(wildfire)
library(tidyverse)
hist(wildfire$fire_size, breaks = 100)
qplot(x = wildfire$fire_size_class, geom = "bar")

wildfire_reduced = data.frame(fire_size,fire_size_class,Vegetation,putout_time,Temp_pre_
30,Temp_pre_15,Temp_pre_7,
                              Temp_cont,Wind_pre_30,Wind_pre_15,Wind_pre_7,Wind_cont,Hum
_pre_30,Hum_pre_15,
                              Hum_pre_7, Hum_cont, Prec_pre_30,Prec_pre_15,Prec_pre_7,Pr
ec_cont, remoteness, latitude, longitude, disc_pre_year, stat_cause_descr)

wildfire_subset = wildfire_reduced[which(wildfire_reduced[1:55367,5] != -1 & wildfire_re
duced[1:55367,13] != 0
                                                            & wildfire_reduced[1:55367,14]
!= 0
                                                            & wildfire_reduced[1:55367,15]
!= 0
                                                            &wildfire_reduced[1:55367,16]
 != 0
                                          & wildfire_reduced[1:55367, 3] != 0),]

for(i in 1:nrow(wildfire_subset)){
  if(wildfire_subset$stat_cause_descr[i] == "Arson"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(1)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Campfire"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(2)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Children"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(3)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Debris Burning"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(4)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Equipment Use"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(5)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Fireworks"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(6)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Lightning"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(7)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Miscellaneous"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(8)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Missing/Undefined"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(9)
  }
```

```
  if(wildfire_subset$stat_cause_descr[i] == "Powerline"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(10)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Railroad"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(11)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Smoking"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(12)
  }
  if(wildfire_subset$stat_cause_descr[i] == "Structure"){
    wildfire_subset$stat_cause_descr[i] = as.numeric(13)
  }
}
#wildfire_subset$Vegetation = as.character(wildfire_subset$Vegetation)
wildfire_subset$stat_cause_descr = as.numeric(wildfire_subset$stat_cause_descr)
#wildfire_subset[1] = wildfire_subset[1]*4046.8564 #convert from acres to square meters.
#NOTE: wildfire_subset is the final subsetted dataset. Use this for modeling and analysi
s.
attach(wildfire_subset)

detach(wildfire)
#detach(wildfire_subset)
subset_F_G = wildfire_subset[which(wildfire_subset$fire_size_class == "F" | wildfire_sub
set$fire_size_class == "G"),]
true.resp = subset_F_G$fire_size_class


FG.test.dat = subset_F_G[c(which(subset_F_G[,2] == "F")[1:223], which(subset_F_G[,2] ==
"G")[1:448]),]
FG.train.dat = subset_F_G[-c(which(subset_F_G[,2] == "F")[1:223], which(subset_F_G[,2] =
= "G")[1:448]),]


for(i in 1:nrow(subset_F_G)){
  if(subset_F_G$fire_size_class[i] == "F"){
    subset_F_G$fire_size_class[i] = 1
  }else{
    subset_F_G$fire_size_class[i] = 0
  }
}

attach(subset_F_G)

detach(wildfire_subset)
library(MASS)
min.model = glm(as.numeric(fire_size_class) ~ 1, data = subset_F_G, family = binomial)

max.model = glm(as.numeric(fire_size_class) ~ ., data = subset.data.frame(subset_F_G, se
lect = -c(fire_size, putout_time)), family = binomial)

step.model = stepAIC(max.model, direction = "both")
```

```
step.model$formula
logmodel = glm(as.numeric(fire_size_class) ~ Temp_pre_30 + Temp_pre_7 + Wind_pre_30 +
    Prec_pre_30 + Prec_pre_15 + remoteness + latitude + longitude +
    disc_pre_year + as.factor(stat_cause_descr), family = binomial, data = subset_F_G)

sm = summary(logmodel)

predictions = predict(logmodel, type = "response")
#predictions

predicted.class = ifelse(predictions > 0.5, "F", "G")

log.model.cm = table(true.resp, predicted.class)
log.model.cm
log.error.rate = 1-sum(diag(log.model.cm))/sum(log.model.cm)
log.error.rate
anova(logmodel, test="Chisq")
logmodel2 = glm(as.numeric(fire_size_class) ~ Temp_pre_7 + Wind_pre_30 +
    Prec_pre_30 + Prec_pre_15 + remoteness + latitude + longitude +
    disc_pre_year + as.factor(stat_cause_descr), family = binomial, data = subset_F_G)

sm2 = summary(logmodel2)

predictions2 = predict(logmodel2, type = "response")
#predictions

predicted.class2 = ifelse(predictions2 > 0.5, "F", "G")
#predicted.class

mean(predicted.class2 == true.resp)
library(class)


FG.knn = knn(train = FG.train.dat[,c(7,9,17,18,21,22,23,24,25)], test = FG.test.dat[,c(7
,9,17,18,21,22,23,24,25)], cl = as.factor(FG.train.dat$fire_size_class), k = 1)

FG.knn.cm = table(FG.test.dat$fire_size_class, FG.knn)
FG.knn.cm
FG.knn.error = 1-sum(diag(FG.knn.cm))/sum(FG.knn.cm)
FG.knn.error
library(MASS)

FG.LDA = lda((fire_size_class) ~ Temp_pre_7 + Wind_pre_30 +
    Prec_pre_30 + Prec_pre_15 + remoteness + latitude + longitude +
    disc_pre_year + as.factor(stat_cause_descr), data = FG.train.dat, family = binomial)

FG.LDA.cm = table(FG.test.dat$fire_size_class, predict(FG.LDA, newdata = FG.test.dat)$cl
ass )
FG.LDA.cm
FG.LDA.error = 1-sum(diag(FG.LDA.cm))/sum(FG.LDA.cm)
FG.LDA.error
tw <- table(subset_F_G$fire_size_class, subset_F_G$stat_cause_descr)
```

```
prob <- tw/sum(tw)
colnames(prob) <- c("Arson", "Campfire", "Children", "Debris Burning", "Equipment Use",
"Fireworks", "Lightning", "Miscellaneous", "Missing/Undefined", "Powerline", "Railroad",
"Smoking", "Structure")
rownames(prob) <- c("Fire_G_", "Fire_F_")
round(prob, digits = 3)
# relevel(factors, ref="")
library(ggplot2)
library("viridis")
par(mfrow=c(2,2))

f.subset = subset_F_G[which(subset_F_G$fire_size_class == 1),]
g.subset = subset_F_G[which(subset_F_G$fire_size_class == 0),]

ggplot(data = f.subset) +
  scale_x_discrete(labels = c("Arson", "Campfire", "Children", "Debris Burning", "Equipm
ent Use", "Fireworks", "Lightning", "Miscellaneous", "Missing/Undefined", "Powerline",
"Railroad", "Smoking", "Structure")) +
  geom_bar(aes(x = as.factor(stat_cause_descr) , fill = fire_size_class)) +
   ggtitle("Causes of Fire by Fire Class Type 'F'") + ylab("Observations of Fire Class
 'F' ") + xlab("Causes of Fire") + scale_fill_brewer(palette = "Set1") + theme_classic()
+ coord_flip()

ggplot(data = g.subset) +
  scale_x_discrete(labels = c("Arson", "Campfire", "Children", "Debris Burning", "Equipm
ent Use", "Fireworks", "Lightning", "Miscellaneous", "Missing/Undefined", "Powerline",
"Railroad", "Smoking", "Structure")) +
  geom_bar(aes(x = as.factor(stat_cause_descr), fill = fire_size_class)) +
  ggtitle("Causes of Fire by Fire Class Type 'G'") + ylab("Observations of Fire Class
 'G' ") + xlab("Causes of Fire") +
  scale_fill_brewer(palette = "Set3") + theme_classic() +  coord_flip()
```