Applied Project

Krishnakanth Kuruvachira Sabu
22078053

MATH7002 Advanced Statistical Method

School of Computer, Data and Mathematical Sciences,
Western Sydney University

Spring, 2024

By including this statement, we the authors of this work, verify that:
• I hold a copy of this assignment that we can produce if the original is lost or damaged.
• I hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
• No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.
• I am aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (which may retain a copy on its database for future plagiarism checking).
• I hereby certify that we have read and understand what the School of Computing and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

```
library(tidyverse)
library(stats)  # For density estimation
library(graphics)
library(mixtools)
require(MASS)
library(pracma)
```

```
# Load the data
fire_data <- read.csv("fire2024.csv")
```

```
# View the first few rows of the dataset
head(fire_data)
```

```
##    latitude longitude      type duration temp humidity
## 1 -34.59785  150.7966    forest       36   28       85
## 2 -35.10680  150.2604    forest       26   23       84
## 3 -34.50271  151.0748 grassland       10   25       52
## 4 -34.13896  150.4528 grassland       16   31       77
## 5 -33.98040  150.0914    forest       49   30       56
## 6 -34.52328  149.8976 grassland       17   24       35
```

```
summary(fire_data)
```

Summary of the data to understand the distributions and possible missing values

```
##     latitude        longitude         type              duration
##  Min.   :-36.19   Min.   :144.4   Length:1959        Min.   : 3.00
##  1st Qu.:-34.22   1st Qu.:149.4   Class :character   1st Qu.:16.00
##  Median :-33.72   Median :150.3   Mode  :character   Median :23.00
##  Mean   :-33.71   Mean   :149.8                      Mean   :25.49
##  3rd Qu.:-33.19   3rd Qu.:150.8                      3rd Qu.:35.00
##  Max.   :-31.17   Max.   :151.9                      Max.   :60.00
```

```
##       temp          humidity
##  Min.   :13.00   Min.   :  1.00
##  1st Qu.:22.00   1st Qu.: 34.00
##  Median :27.00   Median : 53.00
##  Mean   :28.76   Mean   : 52.43
##  3rd Qu.:33.00   3rd Qu.: 74.00
##  Max.   :58.00   Max.   :100.00
```

```
# Check structure of the data
str(fire_data)
```

```
## 'data.frame':    1959 obs. of  6 variables:
##  $ latitude : num  -34.6 -35.1 -34.5 -34.1 -34 ...
##  $ longitude: num  151 150 151 150 150 ...
##  $ type     : chr  "forest" "forest" "grassland" "grassland" ...
##  $ duration : int  36 26 10 16 49 17 11 12 16 35 ...
##  $ temp     : int  28 23 25 31 30 24 20 22 20 29 ...
##  $ humidity : int  85 84 52 77 56 35 30 49 35 56 ...
```

Question 1

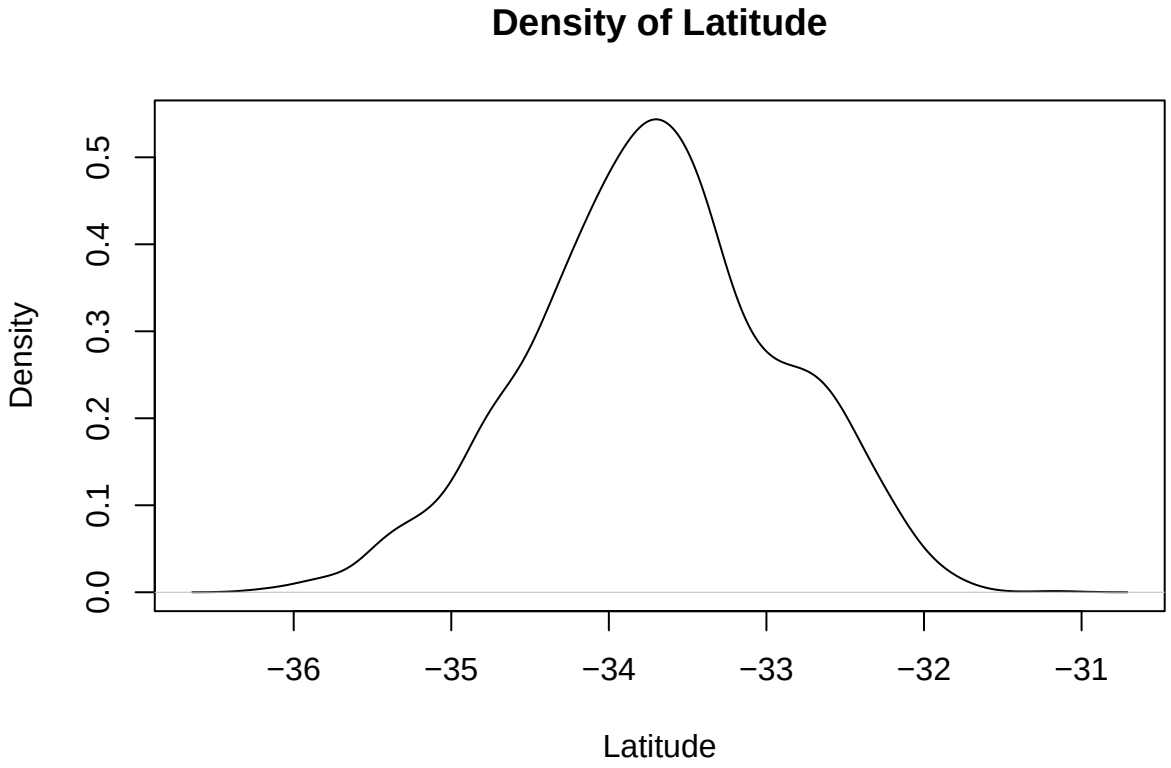
Step 1

Find the density of both latitude and Longitude.

```
# Kernel density estimation for latitude and longitude
density_lat <- density(fire_data$latitude)
density_lon <- density(fire_data$longitude)
```

Plot the Density for Latitude :

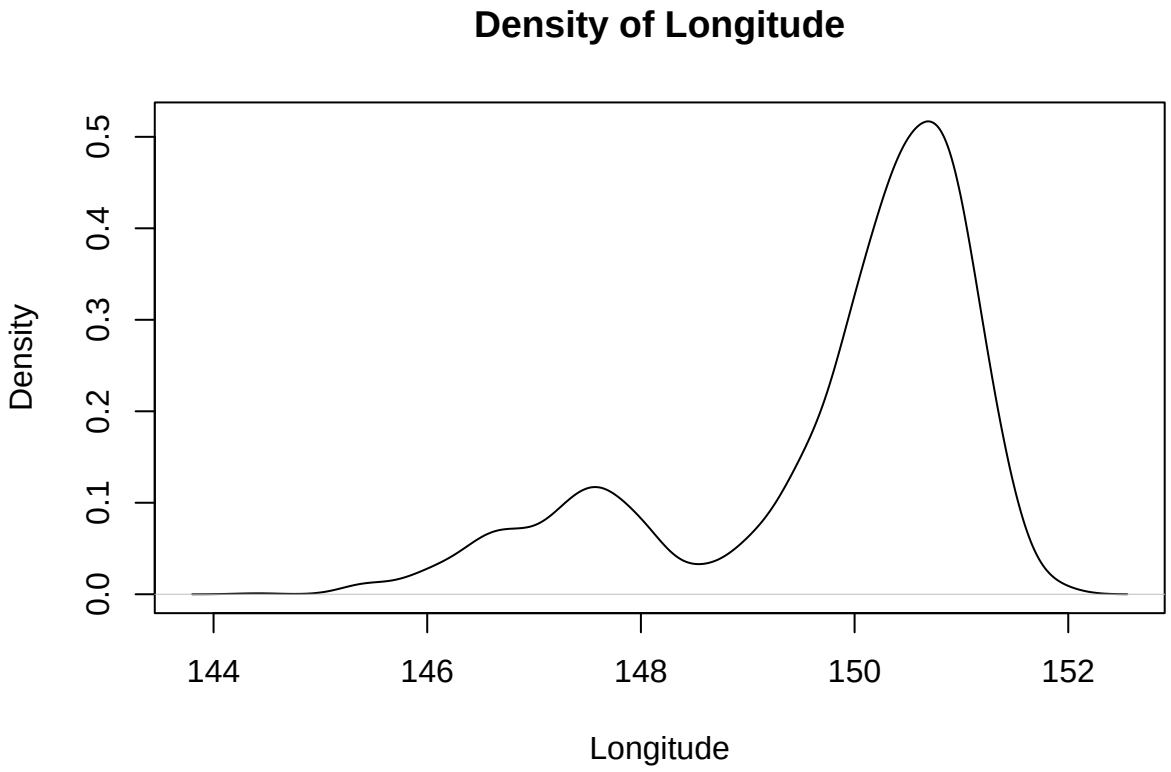
```
# Plotting the densities
plot(density_lat, main = "Density of Latitude", xlab = "Latitude", ylab = "Density")
```



The density plot for latitude reveals that most fires are concentrated between -36° and -32° latitude. This range suggests that the majority of fire incidents occur in the southern regions of NSW.

Plot the Density for Longitude:

```
plot(density_lon, main = "Density of Longitude", xlab = "Longitude", ylab = "Density")
```



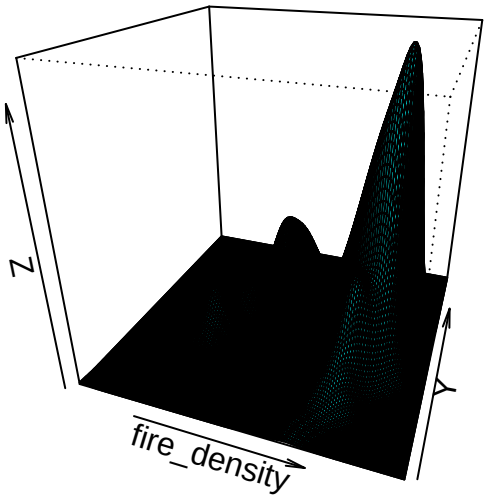
The density plot for longitude shows that most fire occurrences are clustered between 145° and 152° longitude.

Step 2

Using kde2d function used to estimate the density of fire occurrences, we can identify the regions with higher fire densities. we can get a good understanding where fires are more likely to occur and thus where fire retardants should be allocated.

```
fire_density <- kde2d( fire_data$longitude,fire_data$latitude, n = 150)
```

```
persp(fire_density, phi=25, theta=20, col = "cyan")
```

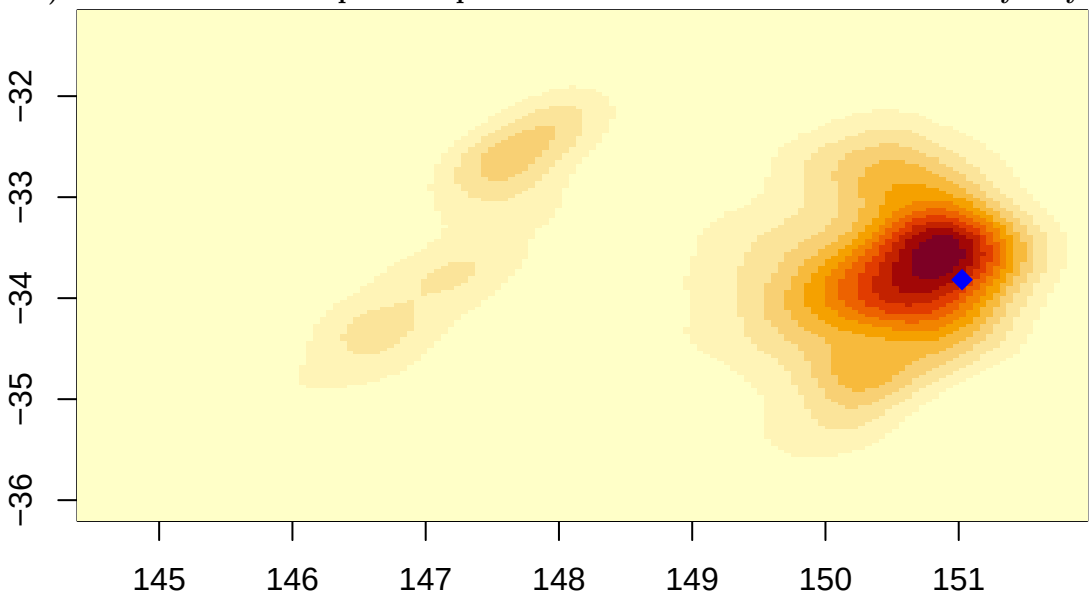


2.1) 3D perspective plot

The peaks and valleys will clearly show which regions are more prone to fires

```
image(fire_density)
points(151.025, -33.817, col = "blue", pch = 18, cex = 1.5)
```

2.2) Plot the heatmap and point the coordinates of western Sydney University.



This heatmap shows fire densities across NSW, with the darker regions indicating higher concentrations of fires. The point in heatmap indicated Western sydney University coordinates.

Step 3

Fit multivariate normal mixture models with different numbers of components (k = 2, 3, 4, 5) to the coordinates. Then calculate the Akaike Information Criterion and Bayesian Information Criterion to compare the models and select the best model.

3.1) Multivariate normal mixture Model Create the matrix with longitudes and latitudes, storing in matrix will make the model run faster.

```
coord_m <- as.matrix(fire_data[, c("longitude", "latitude")])
```

Create all the model with different kernals.

```
mv_normal <- mvnormalmixEM(coord_m, k = 2)
```

```
## number of iterations= 18
```

```
mv_normal1 <- mvnormalmixEM(coord_m, k = 3, maxit = 1000)
```

```
## number of iterations= 417
```

```
mv_normal2 <- mvnormalmixEM(coord_m, k = 4, maxit = 1000)
```

```
## number of iterations= 402
```

```
mv_normal3 <- mvnormalmixEM(coord_m, k = 5, maxit = 1500)
```

```
## number of iterations= 648
```

To model the fire locations across New South Wales (NSW) based on latitude and longitude, a Multivariate Normal Mixture Model was implemented using the function. This model helps identify clusters of fire-prone regions, which can provide useful insights for the Fire Rescue Services.

Step 4

AIC Find the AIC to find the best model among them by choosing the lowest AIC.

```
aic_mv <- c(-2 * mv_normal$loglik + 2 * (6 * 2 -1),
          -2 * mv_normal1$loglik + 2 * (6 * 3 -1),
          -2 * mv_normal2$loglik + 2 * (6 * 4 -1),
          -2 * mv_normal3$loglik + 2 * (6 * 5 -1))
```

```
aic_mv
```

```
## [1] 9587.550 9386.248 9314.716 9290.237
```

AIC : Third Model is better

BIC Find the BIC to find the best model among them by choosing the lowest BIC, for computing .

```
nLen <- nrow(fire_data)
```

```
bic_mv <- c(-2 * mv_normal$loglik + log(nLen) * (6*2-1),
            -2 * mv_normal1$loglik + log(nLen) * (6*2-1),
            -2 * mv_normal2$loglik + log(nLen) * (6*2-1),
            -2 * mv_normal3$loglik + log(nLen) * (6*2-1))
bic_mv
```

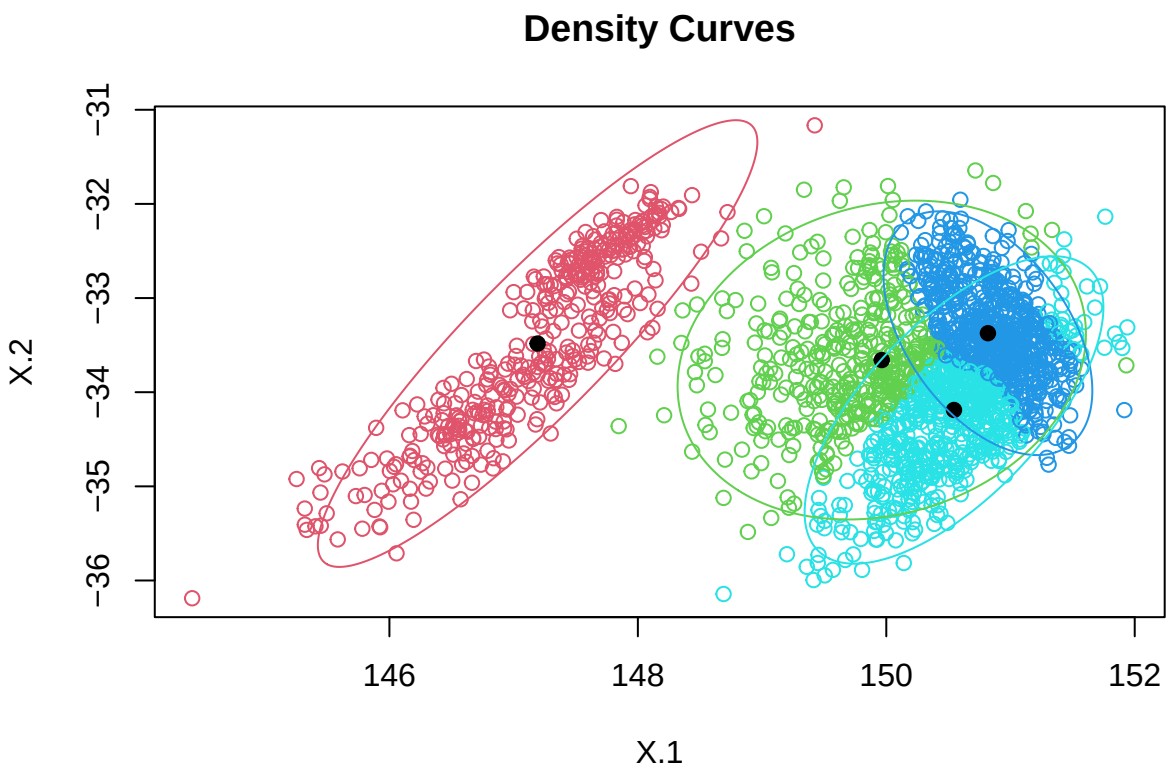
```
## [1] 9648.932 9435.630 9352.098 9315.619
```

BIC is better in : Third model

Step 4.1

Plot the model Best model.

```
plot(mv_normal2, whichplots = 2)
```



The plot shows there are four clusters of fire locations across NSW, which can identify fire alert areas. Each cluster have correspond to regions where fires occur with different frequencies.

Step 4.2

Now after doing the AIC and BIC, we know that model three with kernel 4 is better. Lets take the coefficient by printing the summary of the model.

```
summary(mv_normal2)
```

```
## summary of mvnormalmixEM object:
##             comp 1     comp 2     comp 3     comp 4
## lambda   0.194553   0.274391   0.241327   0.289729
## mu1    147.192990 149.963619 150.818348 150.545184
## mu2    -33.483393 -33.658651 -33.373536 -34.188883
## loglik at estimate:  -4634.358
```

Print and save the coefficient to create the function for the density estimation.

Lambda Value:

```
lambda_model <- mv_normal2$lambda
lambda_model
```

```
## [1] 0.1945527 0.2743908 0.2413275 0.2897290
```

SD:

```
sigma_model <- mv_normal2$sigma
sigma_model
```

```
## [[1]]
##           [,1]      [,2]
## [1,] 0.5224297 0.6321090
## [2,] 0.6321090 0.9394437
##
## [[2]]
##            [,1]       [,2]
## [1,] 0.45088843 0.07660274
## [2,] 0.07660274 0.47863673
##
## [[3]]
##             [,1]        [,2]
## [1,]  0.11786627 -0.08567698
## [2,] -0.08567698  0.27968116
##
## [[4]]
##           [,1]      [,2]
## [1,] 0.2413952 0.2183551
## [2,] 0.2183551 0.4449529
```

Mean:

```
mean_model <- mv_normal2$mu
mean_model
```

```
## [[1]]
## [1] 147.19299 -33.48339
##
## [[2]]
## [1] 149.96362 -33.65865
##
## [[3]]
## [1] 150.81835 -33.37354
##
## [[4]]
## [1] 150.54518 -34.18888
```

Store the coordinates of Western Sydney University to a variable.

```
x <- c(151.0250, -33.8121)
```

Create a density function with all the compound to compute the density estimate of a fire occurring in WSU. Use dmvnorm to create function.

```
func_dens= (lambda_model[[1]] * dmvnorm(x, mu = mean_model[[1]], sigma = sigma_model[[1]])) +
  (lambda_model[[2]] * dmvnorm(x, mu = mean_model[[2]], sigma = sigma_model[[2]])) +
  (lambda_model[[3]]* dmvnorm(x, mu = mean_model[[3]], sigma = sigma_model[[3]])) +
  (lambda_model[[4]] * dmvnorm(x, mu = mean_model[[4]], sigma = sigma_model[[4]]))
```

Call the Function.

```
func_dens
```

```
## [1] 0.305929
```

Step 5

Limitations Lack of Temporal Data: The absence of date information in the data prevented us from incorporating the time of year as a variable in our model.
Model Complexity: As the number of components (clusters) in the mixture model increased, the model became more complex.

Simplified Kernel Assumption: The mixture model assumes multivariate normal distributions for each cluster, which may not fully capture the true distribution of fire occurrences, especially if the fire density patterns do not follow normal distributions.

Step 6 Result and Analysis:

We built a multivariate normal mixture model using latitude and longitude data to estimate fire occurrence densities across NSW. We tested models with different numbers of components (k = 2, 3, 4, and 5) and selected the best one using AIC and BIC criteria. The model with four components (k = 4) provided the best fit, revealing distinct clusters of fire prone areas in NSW.

We then used this model to estimate the density of fire occurrences at Western Sydney University's Parramatta South Campus, with a calculated density of 0.3059. This result helps Fire Rescue Services identify high risk areas and plan fire station locations to ensure better preparedness and resource allocation.

Question 2

Regression Model

Step 1 Build regression mixture with 2, 3, and 4 clusters. arbvar equals to false as we believe that the variance for all the clusters are same.

```
set.seed(10)
regmix_model1 <- with(data = fire_data,(regmixEM(y = duration, x = humidity, k = 2, arbvar = F)))
```

```
## number of iterations= 23
```

```
regmix_model2 <- with(data = fire_data,(regmixEM(y = duration, x = humidity, k = 3, arbvar = F)))
```

```
## number of iterations= 205
```

```
regmix_model3 <- with(data = fire_data,(regmixEM(y = duration, x = humidity, k = 4, arbvar = F)))
```

```
## number of iterations= 269
```

Step 1.1

AIC

Find the AIC to find the best model by choosing the lowest AIC.

```
aic_reg <- c(-2*regmix_model1$loglik+2*(3*2),
        -2*regmix_model2$loglik+2*(3*3),
        -2*regmix_model3$loglik+2*(3*4))
aic_reg
```

```
## [1] 14451.28 14394.40 14361.40
```

Model three is better.

BIC

Find the BIC to find the best model by choosing the best BIC. For computing the BIC take the length of the feature as n

```
n = length(fire_data$duration)
```

```
bic_reg <- c(-2*regmix_model1$loglik+log(n)*(3 * 2),
         -2*regmix_model2$loglik+log(n)*(3 * 3),
         -2*regmix_model3$loglik+log(n)*(3 * 4))
bic_reg
```

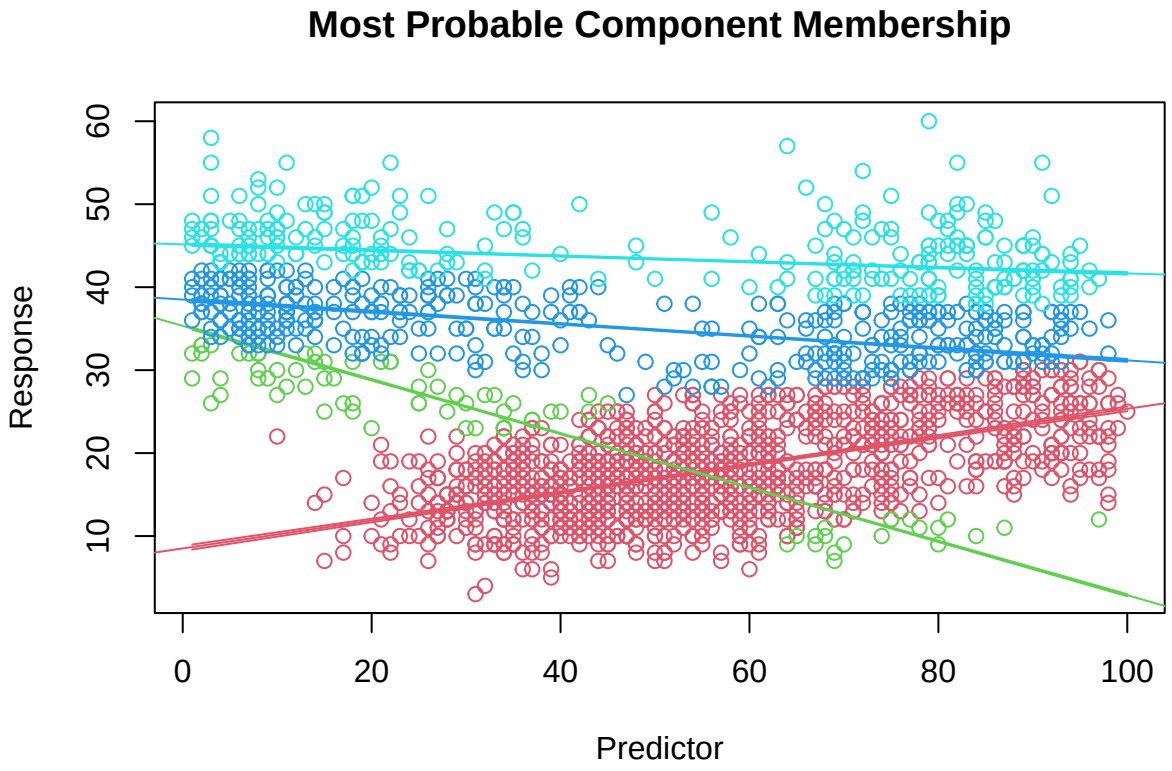
```
## [1] 14484.77 14444.62 14428.36
```

Model three is better.

Step 2

Plot the Best Model

```
plot(regmix_model3, whichplots = 2)
```



This is used when the relationship between a predictor variable and a response variable varies across different subgroups in the data. Each color in the plot represents a different component of the model, and the lines show the regression lines for each of these components.

Step 3

Take the summary of the model.

```r
summary(regmix_model3)
```

```
## summary of regmixEM object:
##          comp 1    comp 2     comp 3     comp 4
## lambda 0.546025  0.110337  0.1994331  0.1442044
## sigma  4.663123  4.663123  4.6631227  4.6631227
## beta1  8.510308 35.331016 38.5056238 45.1568330
## beta2  0.168284 -0.324637 -0.0732869 -0.0349646
## loglik at estimate:  -7168.698
```

```r
regmix_model3$beta
```

```
##           comp.1     comp.2      comp.3      comp.4
## beta.0 8.5103078 35.3310160 38.50562379 45.15683304
## beta.1 0.1682839 -0.3246367 -0.07328693 -0.03496462
```

From this model it is reflected that the relation between the Duration and Humidity.

Step 4

Limitations: Model Complexity: The mixture model increases in complexity as the number of components grows.
Interpret ability: While the mixture model provides more flexibility for different subgroups, interpreting the results becomes more difficult.
Unobserved Variables: Although the mixture model can identify different patterns in the relationship between humidity and fire duration, the exact nature of the hidden factors remains unclear.

Step 5 Results and Analysis: coefficients :
Component 1**:**

Lambda = 0.546025 ; Sigma = 4.663123 ; beta1 which is intercept =8.5103078 ; beta2 which is =0.1682.

Component 2**:**

Lambda = 0.110337 ; Sigma = 4.6631 ; beta1 which is intercept =35.3310 ; beta2 which is = -0.32463.

Component 3:

Lambda = 0.1994331 ; Sigma = 4.6631 ; beta1 which is intercept =38.5056 ; beta2 which is = -0.07328.

Component 4:

Lambda = 0.144204 ; Sigma = 4.6631 ; beta1 which is intercept =45.1568 ; beta2 which is = -0.034964.

Question 3

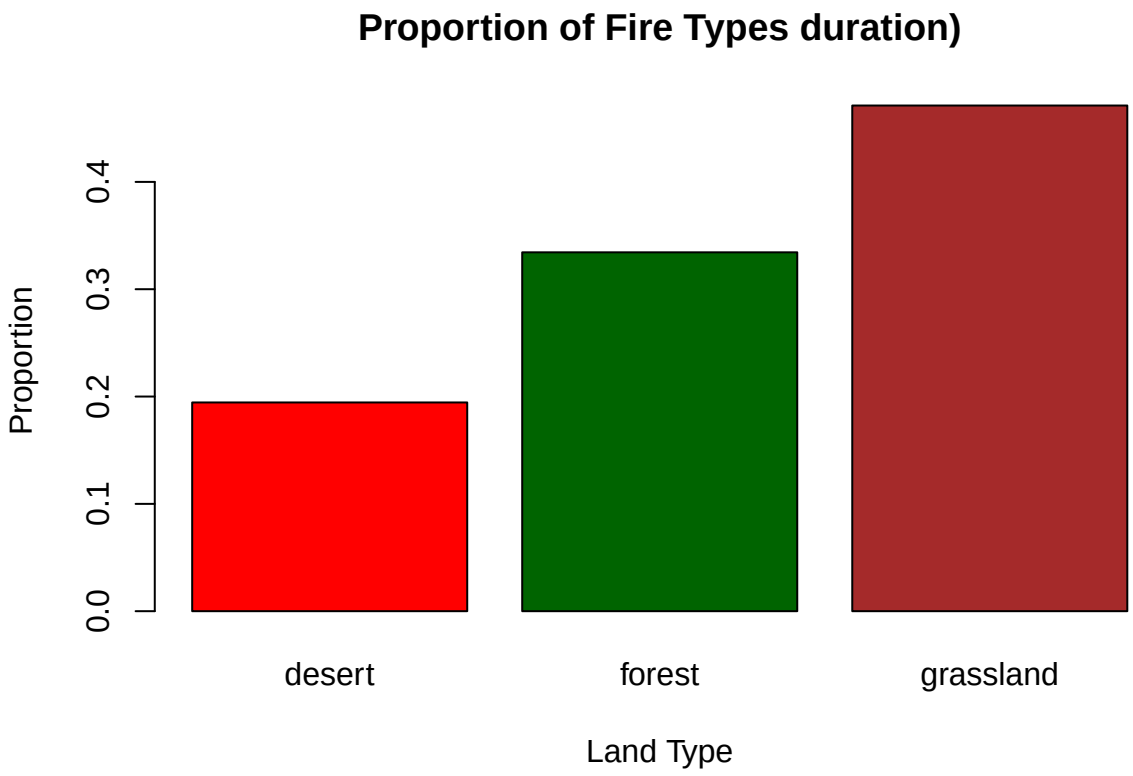
Step 1

Take the types count and plot.

```r
type_count <- table(fire_data$type)
type_probabilities <- prop.table(type_count)
```

We have three types, plot all three types to check the range.

```r
barplot(type_probabilities, main = "Proportion of Fire Types duration)",
        xlab = "Land Type", ylab = "Proportion", col = c("red", "darkgreen", "brown"),
        names.arg = names(type_probabilities))
```



Step 2

Create columns contain to see which row has the particular type.

```r
fire_data$Forest <- ifelse(fire_data$type == "forest", 1, 0)
fire_data$Grassland <- ifelse(fire_data$type == "grassland", 1, 0)
fire_data$Desert <- ifelse(fire_data$type == "desert", 1, 0)  # Assuming 'desert' is a type
```

Check the range of Duration for plotting the density.

```r
range(fire_data$duration)
```

```
## [1]  3 60
```

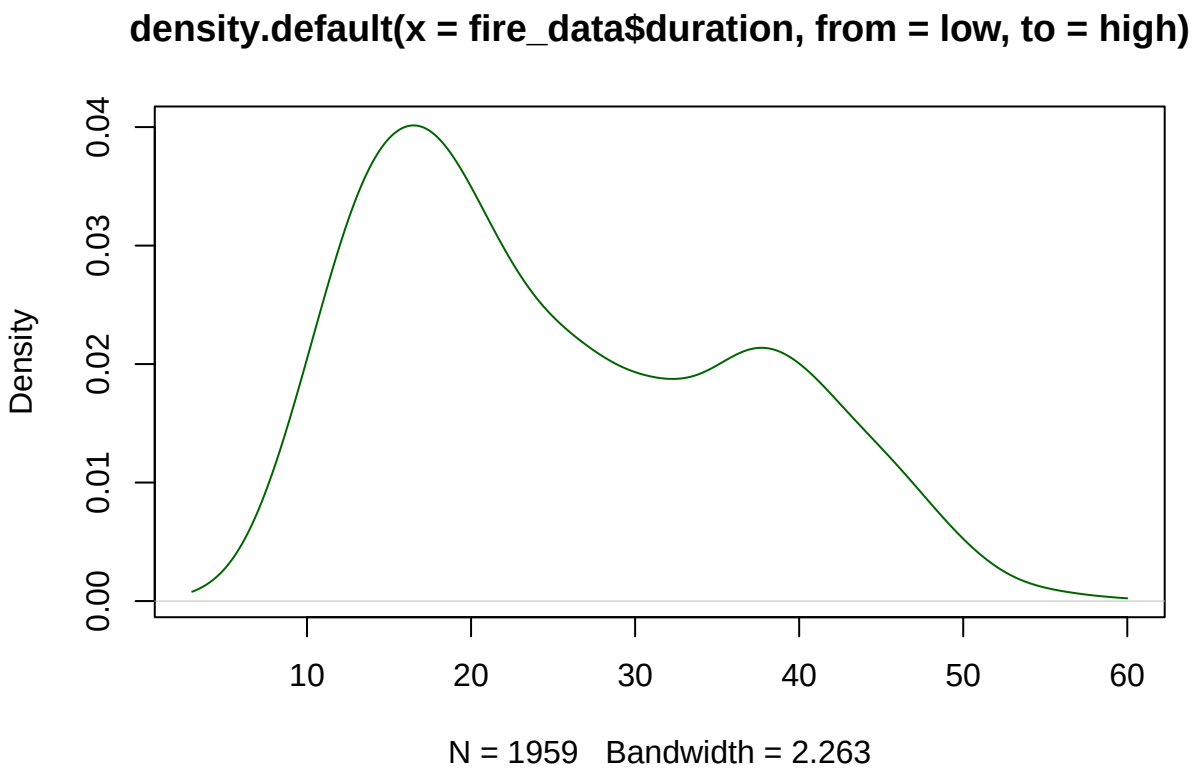
create low and High

```r
low <- 3
high <- 60
```

Step 3

Plot an Unconditional Density graph.

```r
unCond <- density(fire_data$duration, from=low, to=high)
plot(unCond, col = "darkgreen")
```



The density plot reveals that the majority of fires tend to last under 45 minutes, with a notisable range in the 15 to 20 minute range and another in the 35 to 45-minute range.

Step 4

Calculate conditional probabilities for Desert, Greenland, and Forest fires based on duration

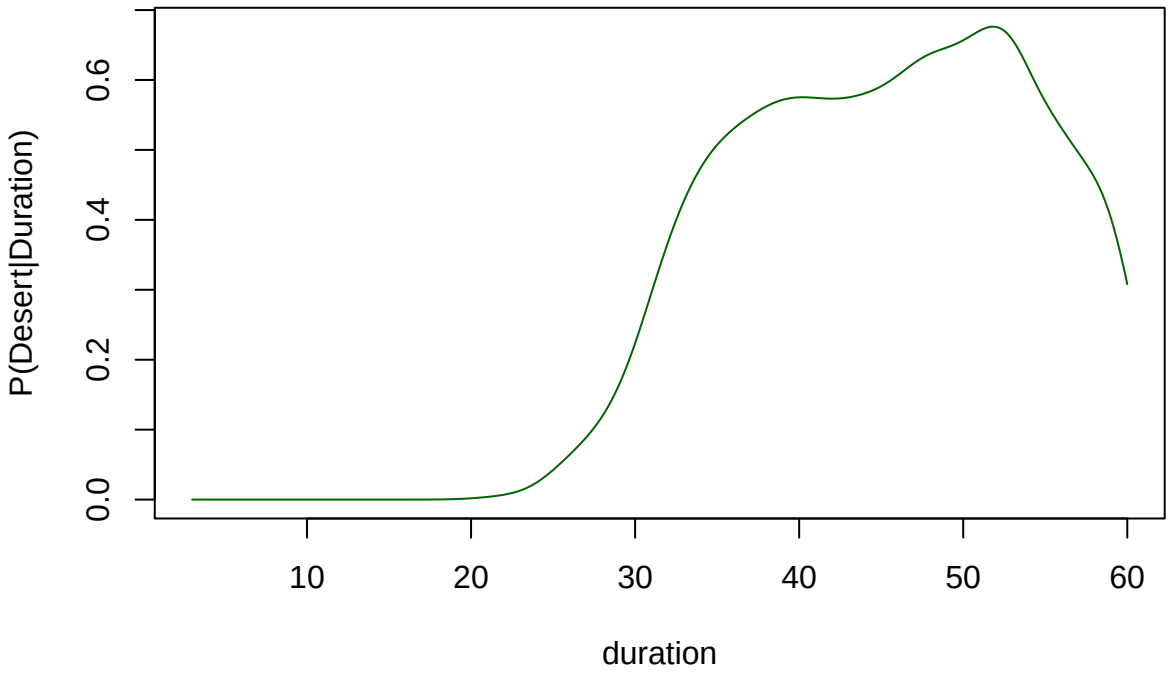
Desert:

```r
prob.desert = sum(fire_data$Desert[fire_data$Desert == 1]) / length(fire_data$duration)
```

```r
f.desert <- density(fire_data$duration[fire_data$Desert==1], from=low, to=high)
```

```r
f.desert.no = density(fire_data$duration[fire_data$Desert==0], from=low, to=high)
```

```r
plot(unCond$x, prob.desert * f.desert$y / (prob.desert*f.desert$y + (1-
prob.desert)*f.desert.no$y), type="l", xlab="duration", ylab="P(Desert|Duration)",
col = "darkgreen")
```



The desert fires are not likely for shorter fire durations but become significantly more probable as the duration increases after 30minutes, peaking around 50 minutes. Fire rescue services should focus their desert fire retardant allocation on fires that are expected to last 35 minutes with the largest proportion reserved for fires that can go beyond 45–50 minutes.

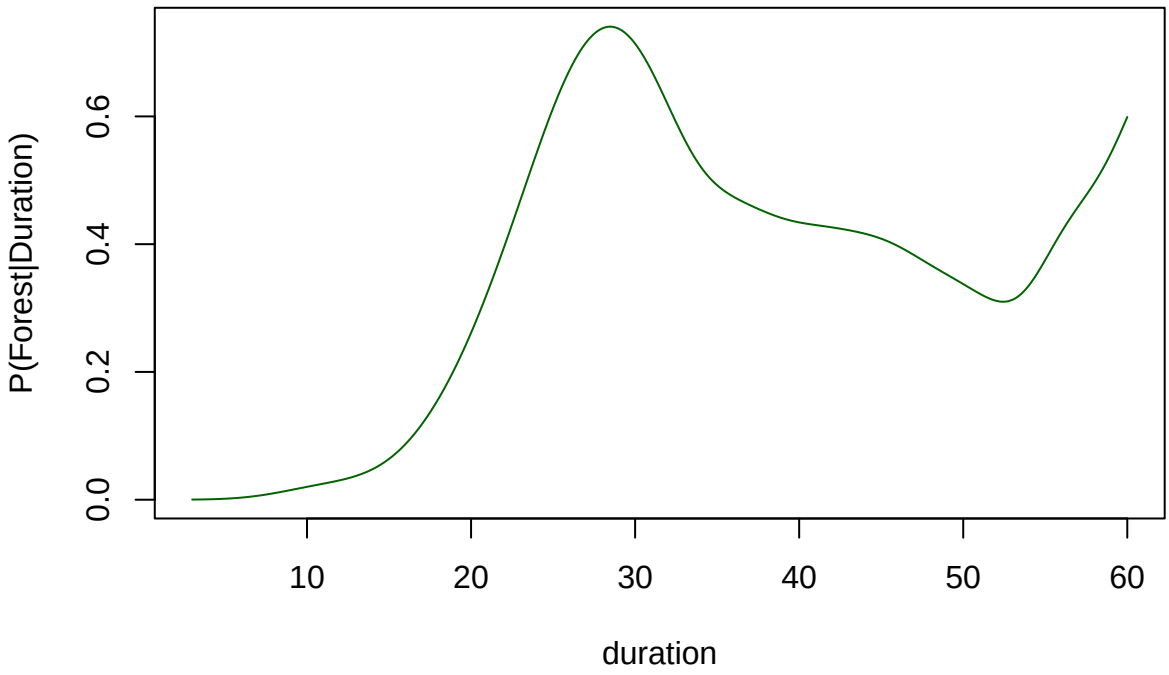
Forest:

```
prob.forest = sum(fire_data$Forest[fire_data$Forest == 1]) / length(fire_data$duration)
```

```
f.forest <- density(fire_data$duration[fire_data$Forest==1], from=low, to=high)
```

```
f.forest.no <- density(fire_data$duration[fire_data$Forest==0], from=low, to=high)
```

```
plot(unCond$x, prob.forest * f.forest$y / (prob.forest*f.forest$y + (1-
prob.forest)*f.forest.no$y), type="l", xlab="duration", ylab="P(Forest|Duration)",
col = "darkgreen")
```



The conditional probability plot for forest fires shows that fires around 30 minutes are likely to occur in forests. For fires lasting 35–45 minutes, which are concern to the fire rescue services, the probability of a forest fire is moderate.

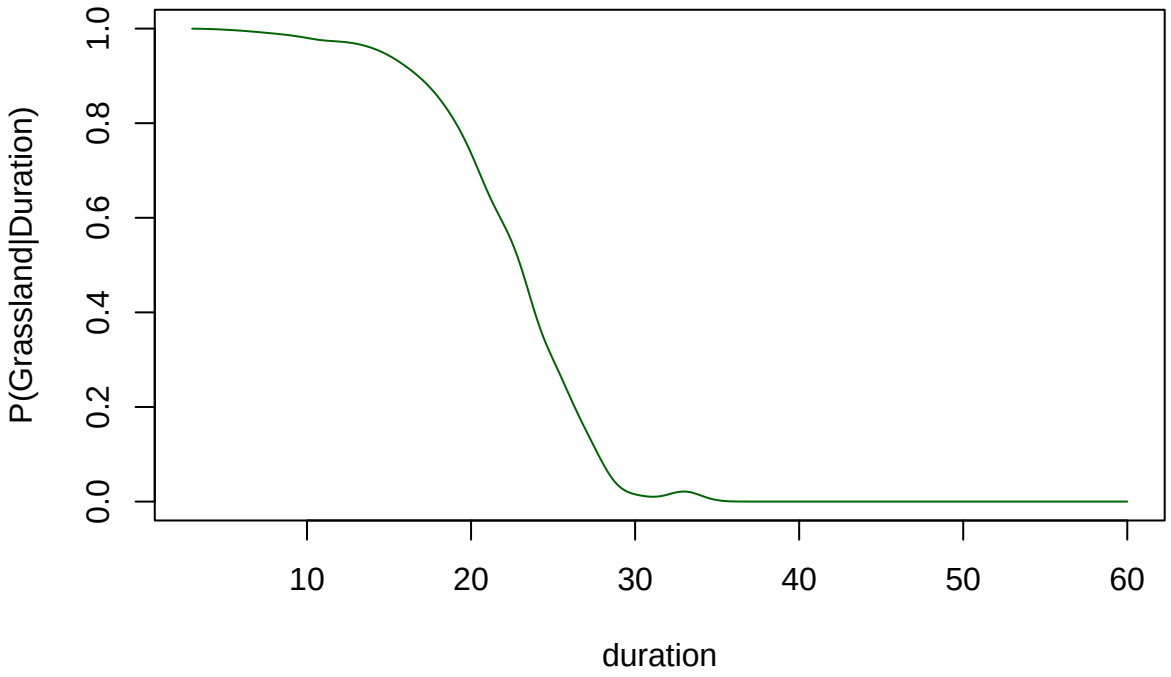
Grass land

```
prob.grass = sum(fire_data$Grassland[fire_data$Grassland == 1]) / length(fire_data$duration)
```

```
f.grass <- density(fire_data$duration[fire_data$Grassland==1], from=low, to=high)
```

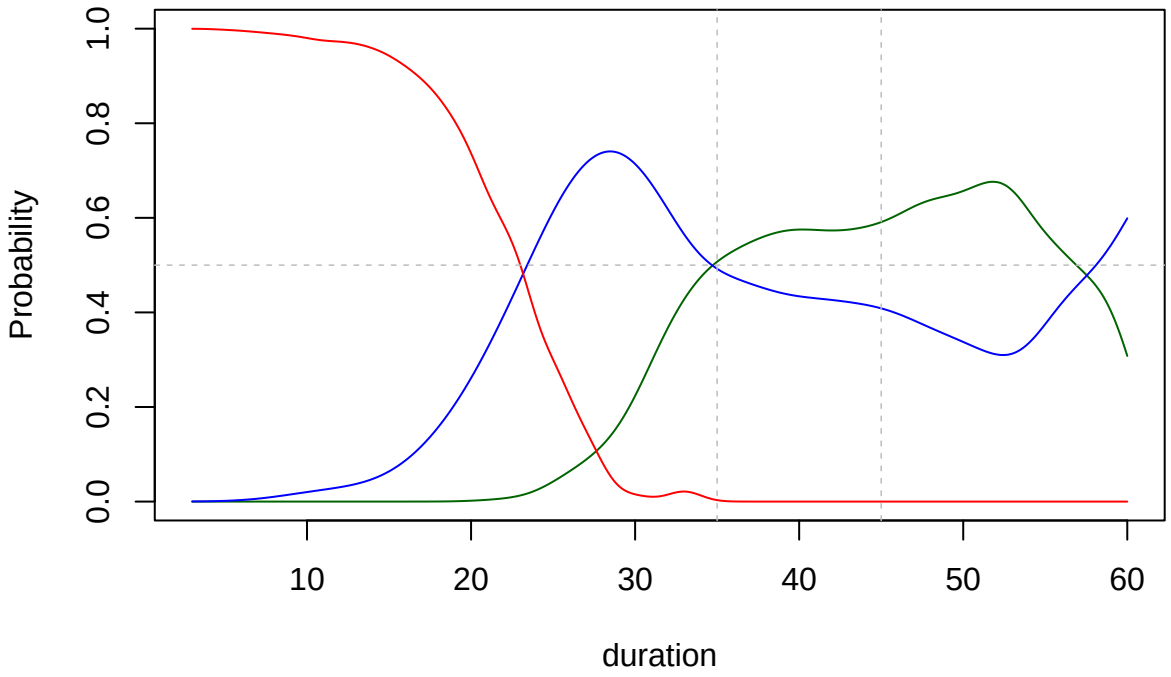
```
f.grass.no <- density(fire_data$duration[fire_data$Grassland==0], from=low, to=high)
```

```
plot(unCond$x, prob.grass * f.grass$y / (prob.grass*f.grass$y + (1-
prob.grass)*f.grass.no$y), type="l", xlab="duration", ylab="P(Grassland|Duration)",
col = "darkgreen")
```



The conditional probability of grassland fires highly drops as duration increases, becoming almost zero for durations over 30 minutes.

```
plot(unCond$x, prob.desert * f.desert$y / (prob.desert*f.desert$y + (1-
prob.desert)*f.desert.no$y), type="l", xlab="duration", ylab="Probability",
col = "darkgreen", ylim = c(0,1))
lines(unCond$x, prob.forest * f.forest$y / (prob.forest*f.forest$y + (1-
prob.forest)*f.forest.no$y), type="l", xlab="duration", ylab="P(Forest|Duration)",
col = "blue")
lines(unCond$x, prob.grass * f.grass$y / (prob.grass*f.grass$y + (1-
prob.grass)*f.grass.no$y), type="l", xlab="duration", ylab="P(Grassland|Duration)",
col = "red")
#abline
abline(v = c(35, 45), h = 0.5,lty = 2, lwd = 0.78, col = "grey")
```



This combined conditional probability plot shows a clear visualization of how the likelihood of different fire types changes with fire duration. The fire rescue services can use this information to make data-driven decisions about how to allocate fire retardants:

- Grassland fire retardants should be focused on short duration fires.
- Forest fire retardants are needed for medium duration fires around 30–35 minutes.
- Desert fire retardants should be reserved for long duration fires 40–50 minutes or more.

Step 5

Analysis of Proportions for Fire Types.

```
duration_filter <-fire_data[fire_data$duration >= 35 & fire_data$duration <= 45,]
```

```
desert_count <- sum(duration_filter$Desert == 1)
forest_count <- sum(duration_filter$Forest == 1)
grass_count <- sum(duration_filter$Grassland == 1)
```

```
duration_range <- nrow(duration_filter)
```

```
proportion_desert <- desert_count / duration_range
proportion_forest <- forest_count / duration_range
proportion_grass <- grass_count / duration_range
```

```
proportion_desert
```

```
## [1] 0.5603865
```

```
proportion_forest
```

```
## [1] 0.4396135
```

```
proportion_grass
```

```
## [1] 0
```

The analysis of the 35–45 minute duration range indicates that desert and forest fires are the primary concerns, with desert fires being a bit more. this allows the fire rescue services to optimize their fire retardant allocation based on the duration of fires, ensuring that resources are efficiently distributed between the fire types most likely to occur within this time frame.

Area Under Curve:

Filter the densities for the 35-45 minute range

```
range_desert <- f.desert$x >= 35 & f.desert$x <= 45
range_forest <- f.forest$x >= 35 & f.forest$x <= 45
range_grass <- f.grass$x >= 35 & f.grass$x <= 45
```

Conditional probabilities using Bayes' theorem.

```
prob_D_range <- prob.desert * f.desert$y[range_desert] /
  (prob.desert * f.desert$y[range_desert] + (1 - prob.desert) * f.desert.no$y[range_desert])

prob_F_range <- prob.forest * f.forest$y[range_forest] /
  (prob.forest * f.forest$y[range_forest] + (1 - prob.forest) * f.forest.no$y[range_forest])

prob_G_range <- prob.grass * f.grass$y[range_grass] /
  (prob.grass * f.grass$y[range_grass] + (1 - prob.grass) * f.grass.no$y[range_grass])
```

```
proportion_desert_auc <- trapz(f.desert$x[range_desert], prob_D_range)
proportion_forest_auc <- trapz(f.forest$x[range_forest], prob_F_range)
proportion_grass_auc <- trapz(f.grass$x[range_grass], prob_G_range)
```

Normalize the areas

```
total_proportion_auc <- proportion_desert_auc + proportion_forest_auc + proportion_grass_auc
proportion_desert_auc <- proportion_desert_auc / total_proportion_auc
proportion_forest_auc <- proportion_forest_auc / total_proportion_auc
proportion_grass_auc <- proportion_grass_auc / total_proportion_auc
```

```
proportion_desert_auc
```

```
## [1] 0.5614463
```

```
proportion_forest_auc
```

```
## [1] 0.4384231
```

```
proportion_grass_auc
```

```
## [1] 0.0001305713
```

Result and Analysis

Both methods the direct proportion calculation and the AUC method converge on result that the fire rescue services should prioritize desert and forest fire retardants, for fires expected to last between 35 and 45 minutes :

56.1% of resources should be allocated to desert fires.

43.8% should be allocated to forest fires.

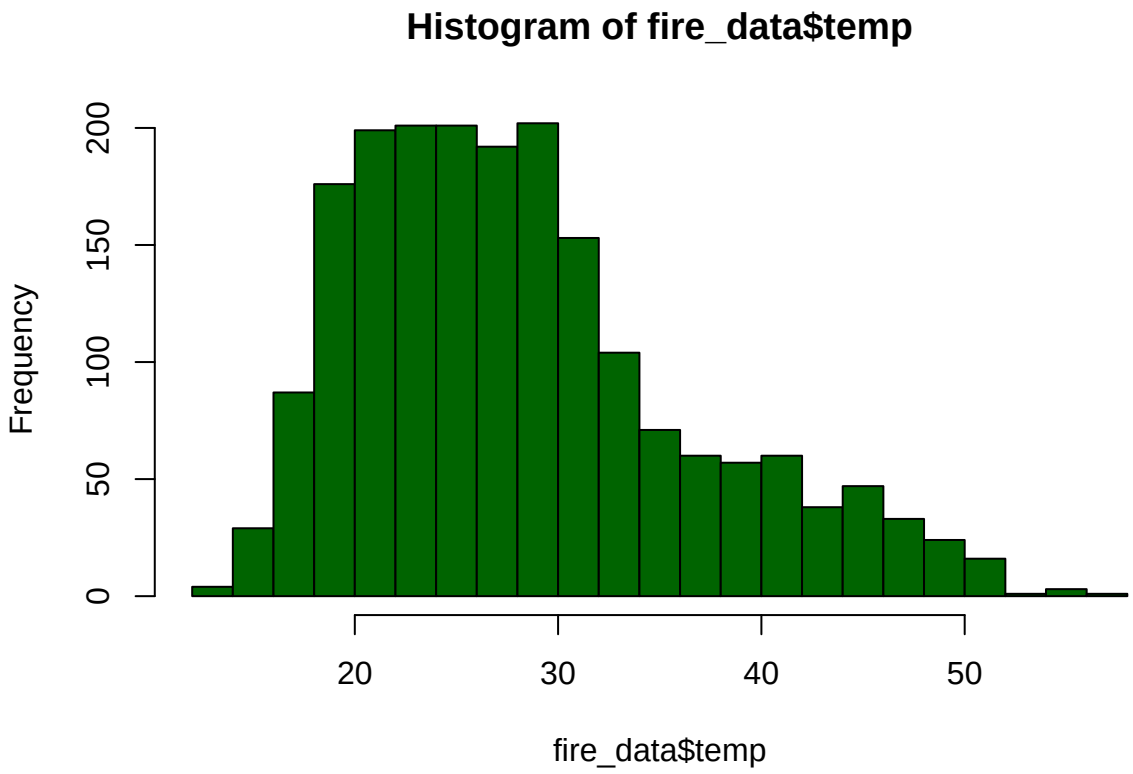
0.1% should be allocated to grassland fires, as their probability of occurring within this duration range is neraly zero.

By utilizing both approaches, we confirm that desert and forest fires dominate this duration range, with desert fires being slightly more common. The alignment of results from both methods provides a strong basis for decision making in the allocation of fire retardants, ensuring that resources are distributed efficiently based on the most likely fire types.

Question 4

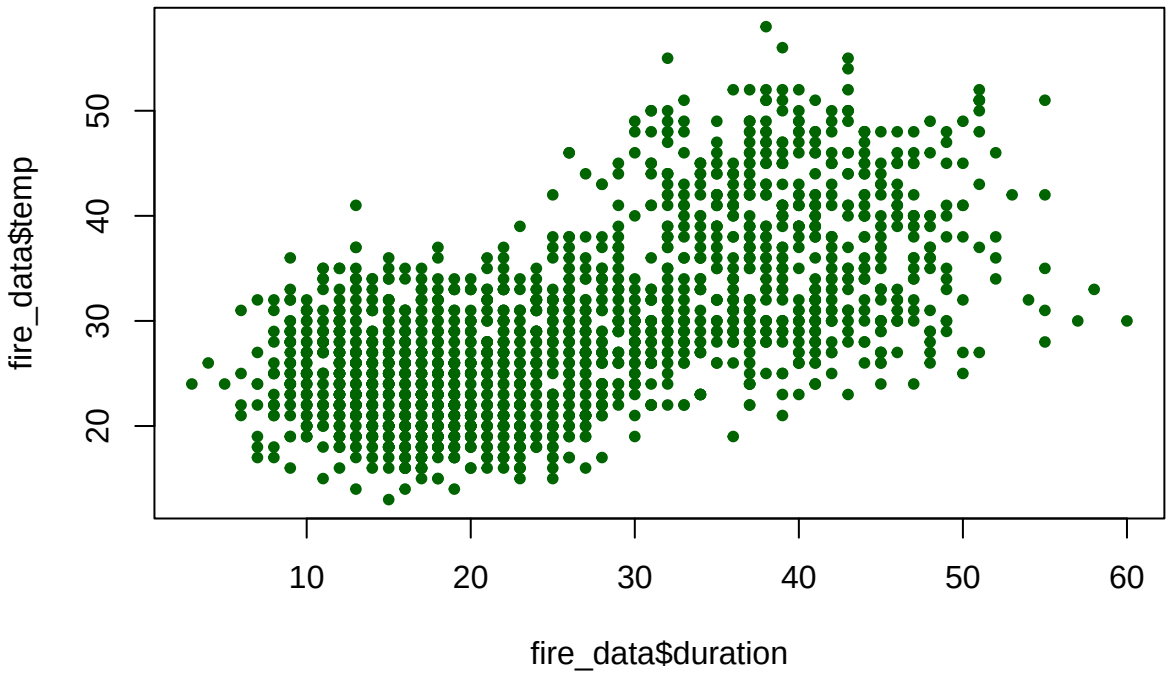
Step 1

```
hist(fire_data$temp, breaks = 20, col = "darkgreen")
```



The histogram of temperature suggests that the data may contain multiple clusters (modes), which could correspond to seasonal effects.

```
plot(fire_data$duration,fire_data$temp, pch = 20, col = "darkgreen")
```



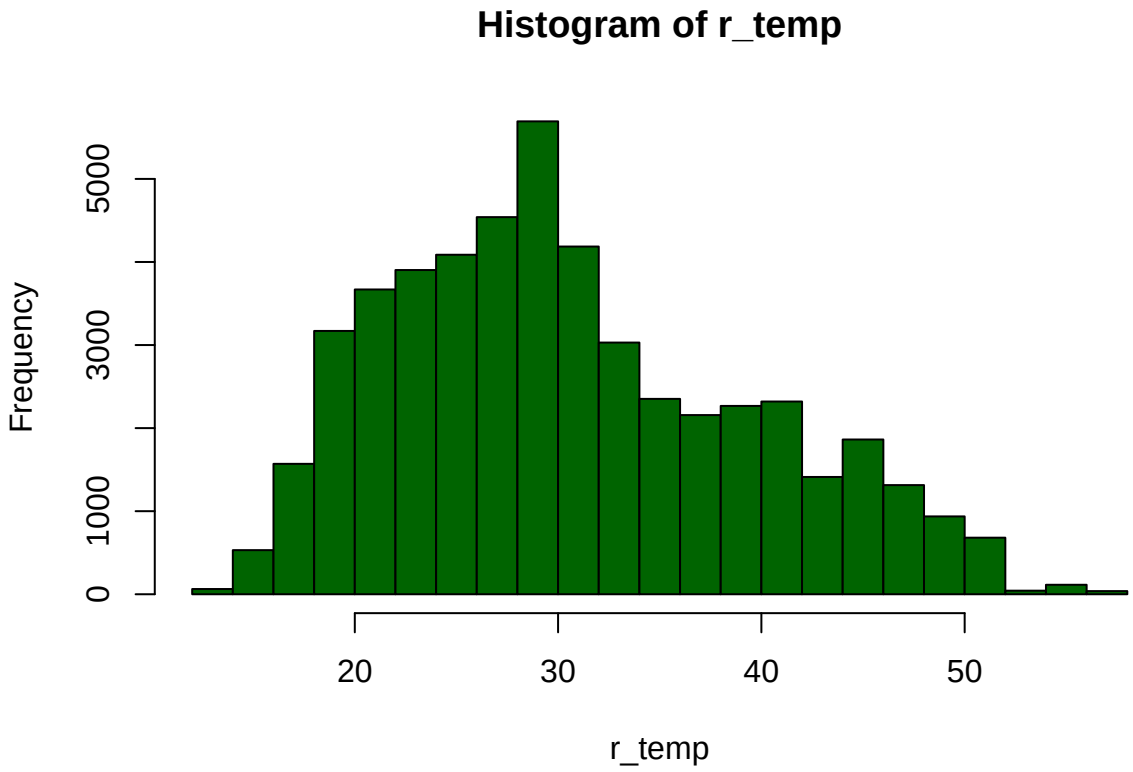
The scatter plot between fire duration and temperature indicates that higher temperatures are generally associated with longer fire duration, which may shows that warmer seasons like summer.

Step 2

```
r_temp <- rep(fire_data$temp, fire_data$duration)
```

By repeating the temperature values on the duration, we are giving more weight to temperatures associated with fires This helps us show how often specific temperatures are experienced.

```
hist(r_temp, col = "darkgreen")
```



This histogram shows a more higten right skewed distribution, with peak at around 30°C. This indicates that fires that occur at this temperature range are common and likely to last longer.

Step 3

Fit Normal Mix for 5 kernals.

```
normal_em <- normalmixEM(r_temp, k = 2)
```

```
## number of iterations= 287
```

```
normal_em2 <- normalmixEM(r_temp, k = 3)
```

```
## number of iterations= 644
```

```
normal_em3 <- normalmixEM(r_temp, k = 4,  maxit = 2000, lambda = NULL)
```

```
## number of iterations= 1356
```

```
normal_em4 <- normalmixEM(r_temp, k = 5, maxit = 3000, lambda = NULL)
```

```
## One of the variances is going to zero;  trying new starting values.
## One of the variances is going to zero;  trying new starting values.
## One of the variances is going to zero;  trying new starting values.
## number of iterations= 1479
```

AIC

```
aic2 <- c(-2*normal_em$loglik+2*(3*2-1),
        -2*normal_em2$loglik+2*(3*3-1),
        -2*normal_em3$loglik+2*(3*4-1),
        -2*normal_em4$loglik+2*(3*5-1))
```

```
aic2
```

```
## [1] 352942.3 351321.1 351098.8 351028.7
```

Model four is better.

BIC

```
n2 <- length(r_temp)
```

```
bic2 <- c(-2*normal_em$loglik+log(n2)*(3 * 2 - 1),
          -2*normal_em2$loglik+log(n2)*(3 * 3 - 1),
          -2*normal_em3$loglik+log(n2)*(3 * 4 - 1),
          -2*normal_em4$loglik+log(n2)*(3 * 5 - 1))
```

```
bic2
```

```
## [1] 352986.4 351391.6 351195.8 351152.2
```

BIC : Model 4 is the Best.

Step 4

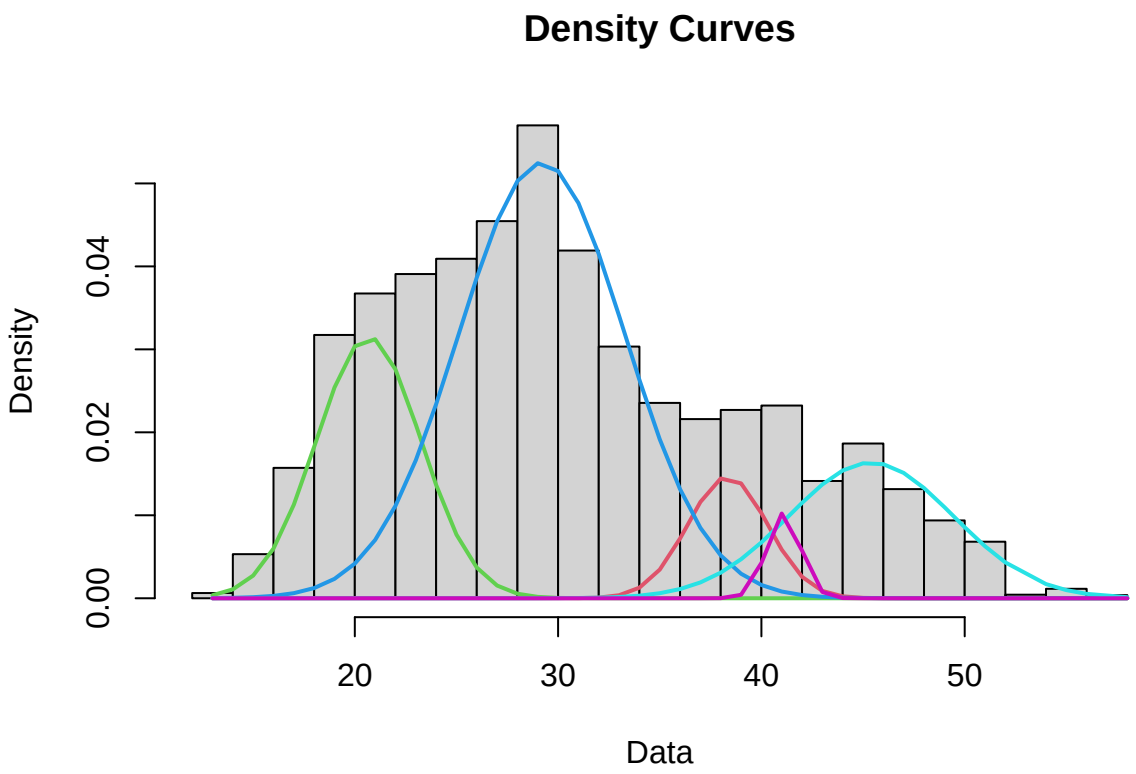
Take summary of the model and plot it.

```
summary(normal_em4)
```

```
## summary of normalmixEM object:
##            comp 1    comp 2    comp 3    comp 4     comp 5
## lambda  0.0721573  0.202693  0.537819  0.165919  0.0214109
## mu     38.3407850 20.681421 29.197023 45.394064 41.1034022
## sigma   1.9646291  2.570059  4.087110  4.050316  0.8298917
## loglik at estimate:  -175500.4
```

We fitted a 5 component Mixture Model as the best model to the temperature data, as this model provides the best fit based on both AIC and BIC. Each component of the model represents a distribution that captures a distinct cluster of temperatures. By examining the characteristics of these components which are means, variances, and mixing proportions, we can interpret how temperature, and therefore seasonality, is distributed across the data.

```
plot(normal_em4, whichplots = 2, col1 = "red")
```



Step 5

Result and Analysis.

By grouping the data into five temperature clusters, the model shows distinct seasonal patterns ranging from cooler winter months 19°C to extreme summer heat 45°C. The largest component, centered around 29°C, suggests that most fires occur during mild temperatures, and in spring or early summer. Meanwhile, extreme summer conditions are represented by the smaller clusters with high temperatures, which may correspond to longer intense fires.