```r
library(mixtools)
```

```
## mixtools package, version 2.0.0, Released 2022-12-04
## This package is based upon work supported by the National Science Foundation under Grant No. SES-0518
```

Cat Breeds

The RCSS want to investigate which times of the year have the most cats births. Use the data to provide a density plot of cat births throughout the year. Examine the plot and provide an appropriate model of the density plot and its fitted coefficients. Present the equation of the model in your report.

```r
data <- read.csv("cats2023.csv")
head(data)
```

```
##    latitude longitude  breed income litter_size gestation       date
## 1 -32.31198  150.4861 Bengal  75992           1        66 2022-09-22
## 2 -32.48031  150.7269 Bengal  74248           2        65 2022-09-23
## 3 -32.21688  150.2134 Bengal  70268           8        60 2022-03-16
## 4 -33.11854  150.3796 Bengal  69344           3        59 2022-07-12
## 5 -32.58464  150.9601 Bengal  70147           6        60 2022-03-25
## 6 -31.93354  150.5808 Bengal  65336           3        55 2022-05-30
```

```r
str(data)
```

```
## 'data.frame':    805 obs. of  7 variables:
##  $ latitude   : num  -32.3 -32.5 -32.2 -33.1 -32.6 ...
##  $ longitude  : num  150 151 150 150 151 ...
##  $ breed      : chr  "Bengal" "Bengal" "Bengal" "Bengal" ...
##  $ income     : int  75992 74248 70268 69344 70147 65336 73819 71667 74970 66666 ...
##  $ litter_size: int  1 2 8 3 6 3 7 8 4 6 ...
##  $ gestation  : int  66 65 60 59 60 55 63 61 65 57 ...
##  $ date       : chr  "2022-09-22" "2022-09-23" "2022-03-16" "2022-07-12" ...
```

```r
summary(data)
```

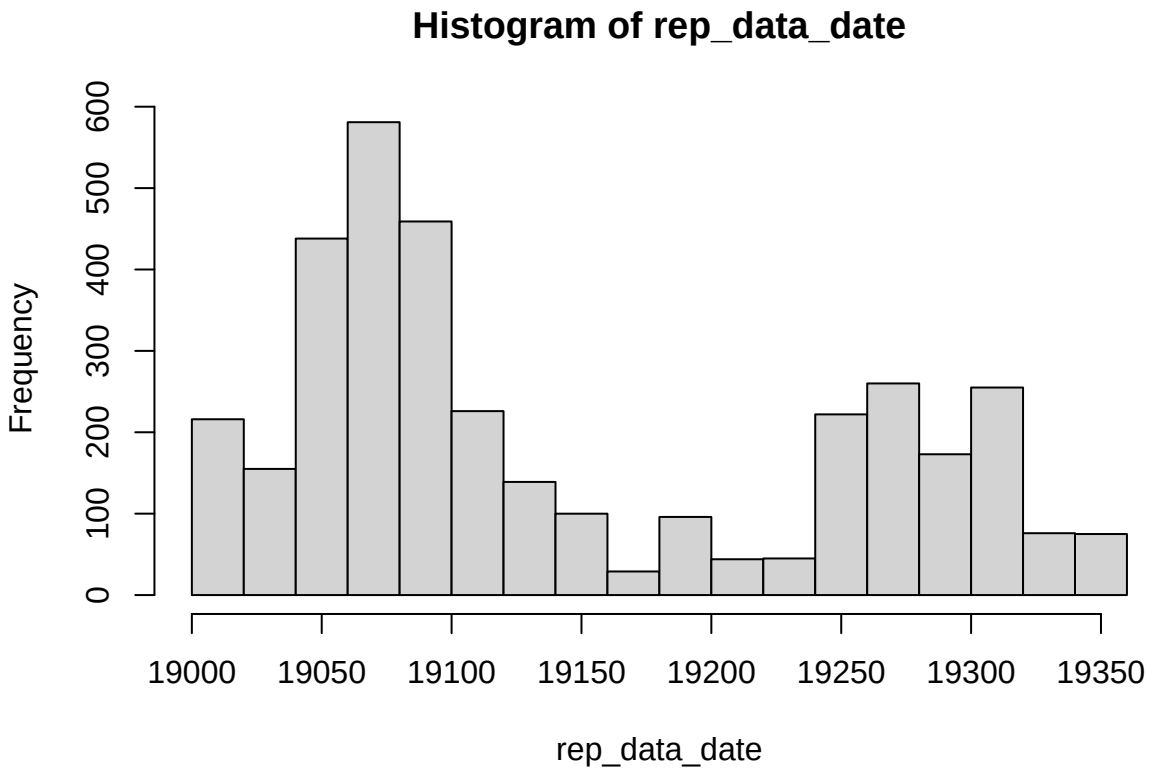
```
##     latitude        longitude        breed               income
##  Min.   :-36.17   Min.   :147.3   Length:805         Min.   : 55573
##  1st Qu.:-34.30   1st Qu.:149.8   Class :character   1st Qu.: 68179
##  Median :-33.65   Median :150.3   Mode  :character   Median : 72203
##  Mean   :-33.65   Mean   :150.2                      Mean   : 84342
##  3rd Qu.:-32.94   3rd Qu.:150.7                      3rd Qu.: 79220
##  Max.   :-31.45   Max.   :152.3                      Max.   :163839
##   litter_size       gestation         date
##  Min.   : 0.000   Min.   :45.00   Length:805
##  1st Qu.: 3.000   1st Qu.:58.00   Class :character
##  Median : 4.000   Median :62.00   Mode  :character
##  Mean   : 4.458   Mean   :61.96
##  3rd Qu.: 6.000   3rd Qu.:66.00
##  Max.   :12.000   Max.   :79.00
```

```r
dim(data)
```

```
## [1] 805   7
```

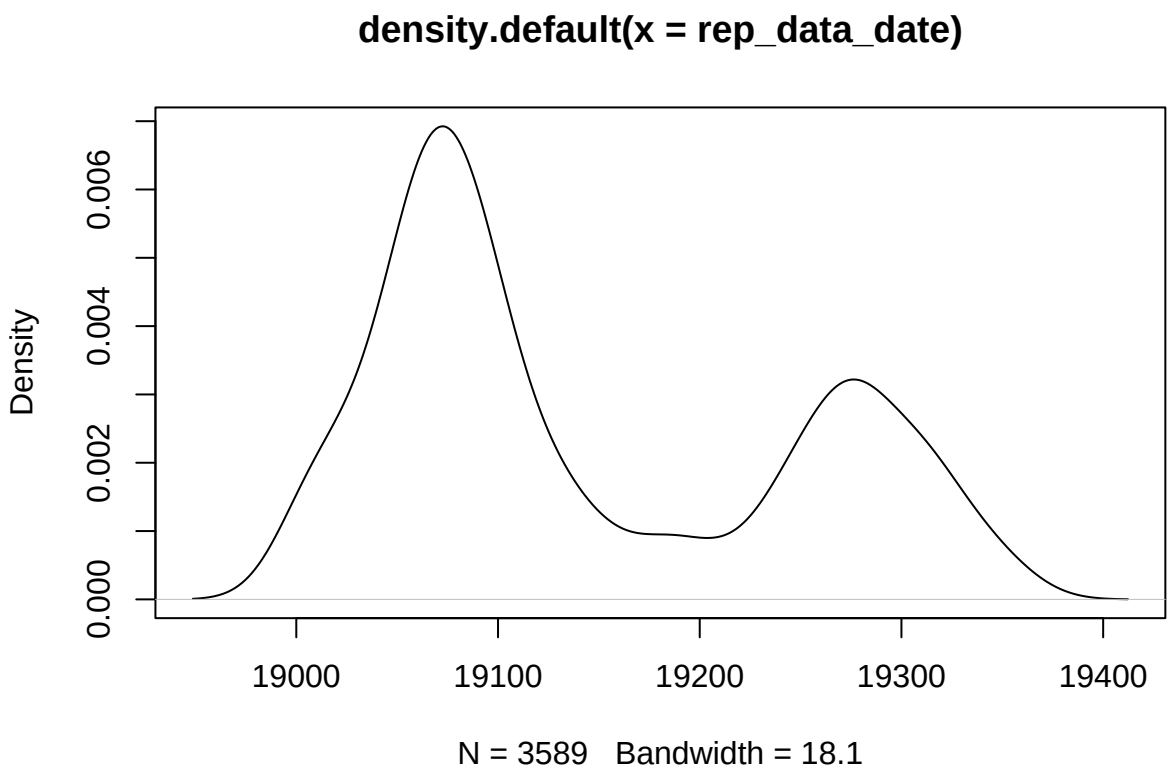
```r
data$date <- as.Date(data$date)
```

```r
rep_data_date <- rep(data$date, data$litter_size)
rep_data_date <- as.numeric(rep_data_date)
hist(rep_data_date)
```



In the histogram above we can see a distribution that has two peaks, so we can suggest that this is a bimodal distribution.

```r
density_model <- density(rep_data_date)
plot(density_model)
```




```
as.Date(density_model$x[which.max(density_model$y)], origin = "1970-01-01")
```

```
## [1] "2022-03-21"
```

From the density plot of the cats' births above we can observe that most cats are born during March-April and October-November.

```
x = rep_data_date

negloglik2 = function(mu1,s1, mu2,s2, lambda)
  -sum(log(lambda*dnorm(x, mu1, s1) + (1-lambda)*dnorm(x,mu2,s2)))

require(stats4)
```

```
## Loading required package: stats4
```

```
fit0 = mle(negloglik2,
    start=list(mu1=min(x), s1=sd(x), mu2=max(x), s2=sd(x), lambda=0.5))

summary(fit0)
```

```
## Maximum likelihood estimation
##
## Call:
## mle(minuslogl = negloglik2, start = list(mu1 = min(x), s1 = sd(x),
##     mu2 = max(x), s2 = sd(x), lambda = 0.5))
##
## Coefficients:
```

```
##            Estimate Std. Error
## mu1    1.907316e+04  0.8376551
## s1     3.660015e+01  0.6888999
## mu2    1.927369e+04  1.4169400
## s2     4.358918e+01  1.1333206
## lambda 6.503564e-01  0.0083054
##
## -2 log L: 40936.18
```

Equation of the model:

f(x) = lambda * exp(-(x - mu1)^2 / (2 * s1^2)) + (1 - lambda) * exp(-(x - mu2)^2 / (2 * s2^2))

where:

- x is the value of date
- lambda is the mixing parameter
- mu1 and mu2 are the means of the two normal distributions
- s1 and s2 are the standard deviations of the two normal distributions

Breed vs Family Income

Certain breeds of cat seem to be chosen to represent social status, we want to investigate if the data shows this relationship. Estimate the probability of each cat breed conditioned on the family income. Provide estimates of the proportion of each cat breed, given that the family income is $80,000.

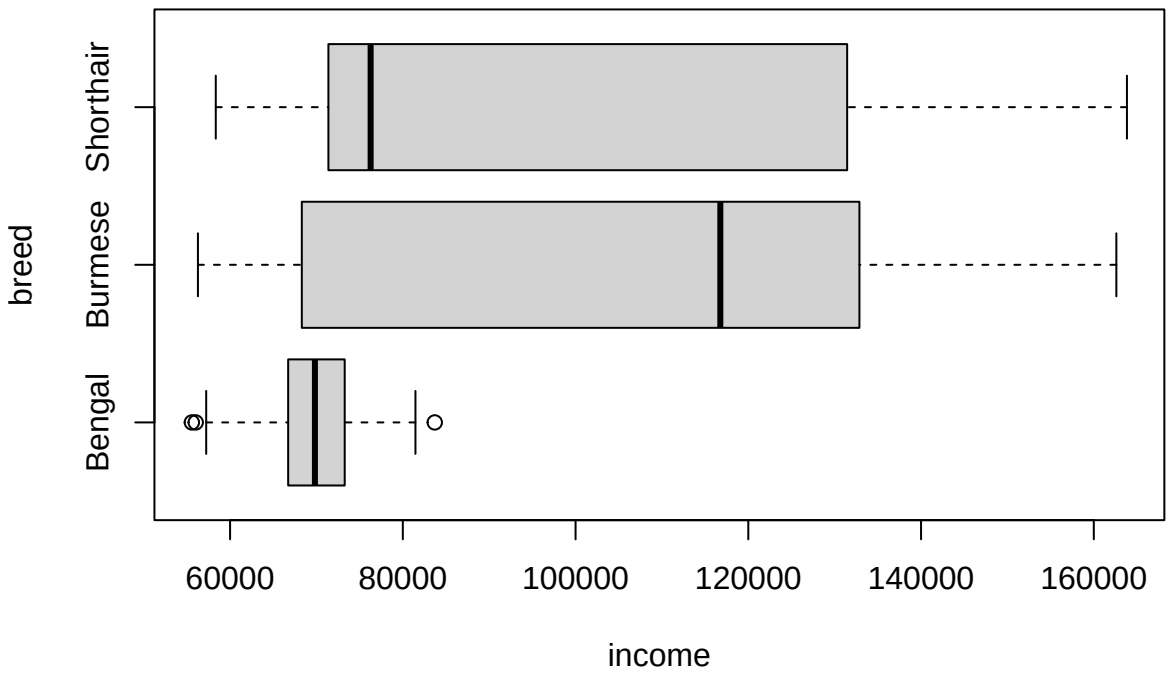
```r
range(data$income)
```

```
## [1]  55573 163839
```

```r
unique(data$breed)
```

```
## [1] "Bengal"    "Shorthair" "Burmese"
```

```r
boxplot(income ~ breed, data = data, horizontal = TRUE)
```



The boxplot shows that the median income for Shorthair cats is the highest, followed by Burmese cats and then Bengal cats. The IQR for Shorthair cats is also the smallest, followed by Burmese cats and then Bengal cats. This means that the income for Shorthair cats is more tightly clustered around the median than the income for Burmese cats or Bengal cats.

The boxplot also shows that there are no outliers in the data for Shorthair cats or Burmese cats. However, there are a few outliers in the data for Bengal cats. This means that there are a few Bengal cats that have incomes that are much higher or lower than the median income for Bengal cats.

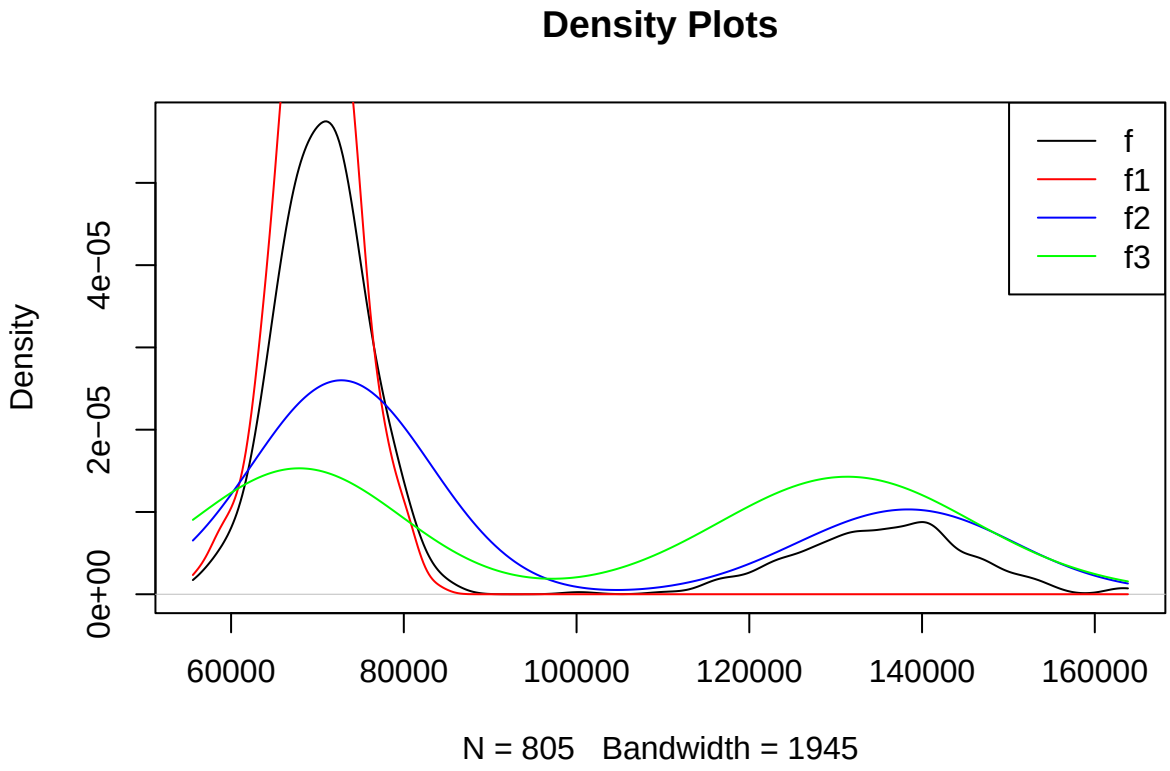
```
lo <- range(data$income)[1]
hi <- range(data$income)[2]

f <- density(data$income, from = lo, to = hi)
f1 <- density(data$income[data$breed == unique(data$breed)[1]], from = lo, to = hi) #Bengal
f2 <- density(data$income[data$breed == unique(data$breed)[2]], from = lo, to = hi) #Shorthair
f3 <- density(data$income[data$breed == unique(data$breed)[3]], from = lo, to = hi) #Burmese
```

```
plot(f, main="Density Plots")

lines(f1, col="red")
lines(f2, col="blue")
lines(f3, col="green")

legend("topright", legend=c("f", "f1", "f2", "f3"), col=c("black", "red", "blue", "green"), lty=1)
```




```
p1 <- mean(data$breed == unique(data$breed)[1]) #Bengal
p1
```

```
## [1] 0.4596273
```

```
p2 <- mean(data$breed == unique(data$breed)[2]) #Shorthair
p2
```

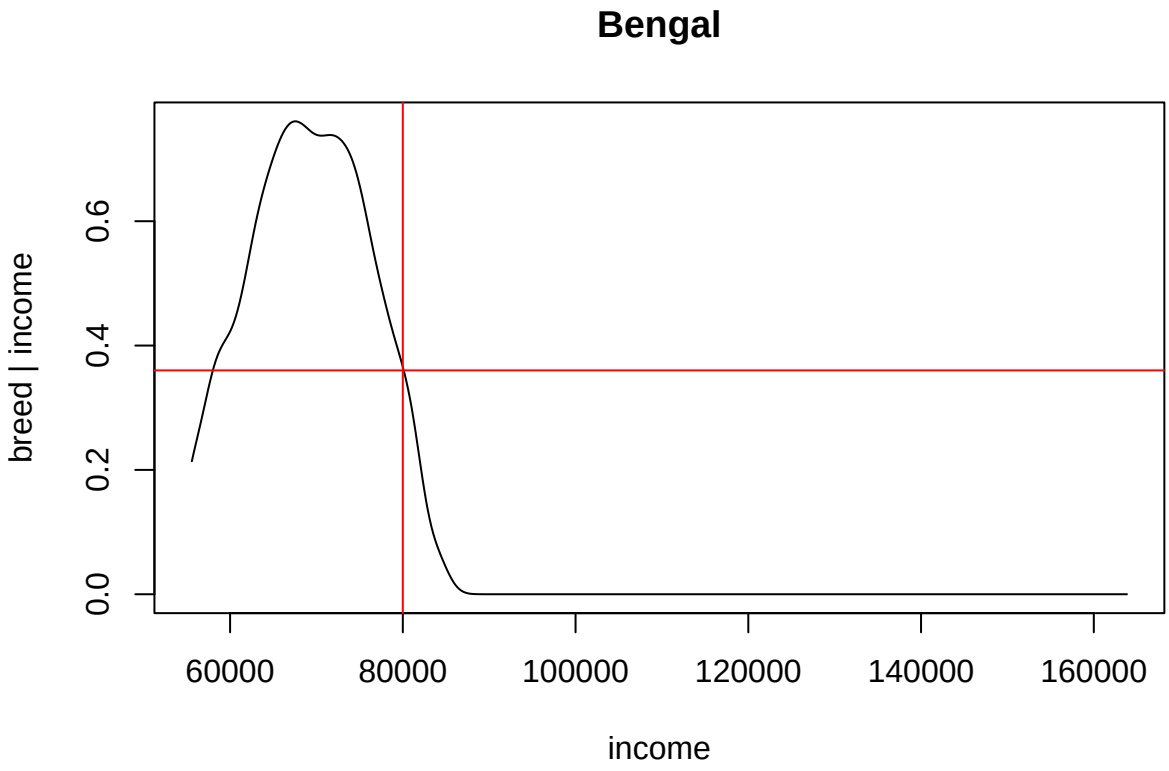
```
## [1] 0.3726708
```

```
p3 <- mean(data$breed == unique(data$breed)[3]) #Burmese
p3
```

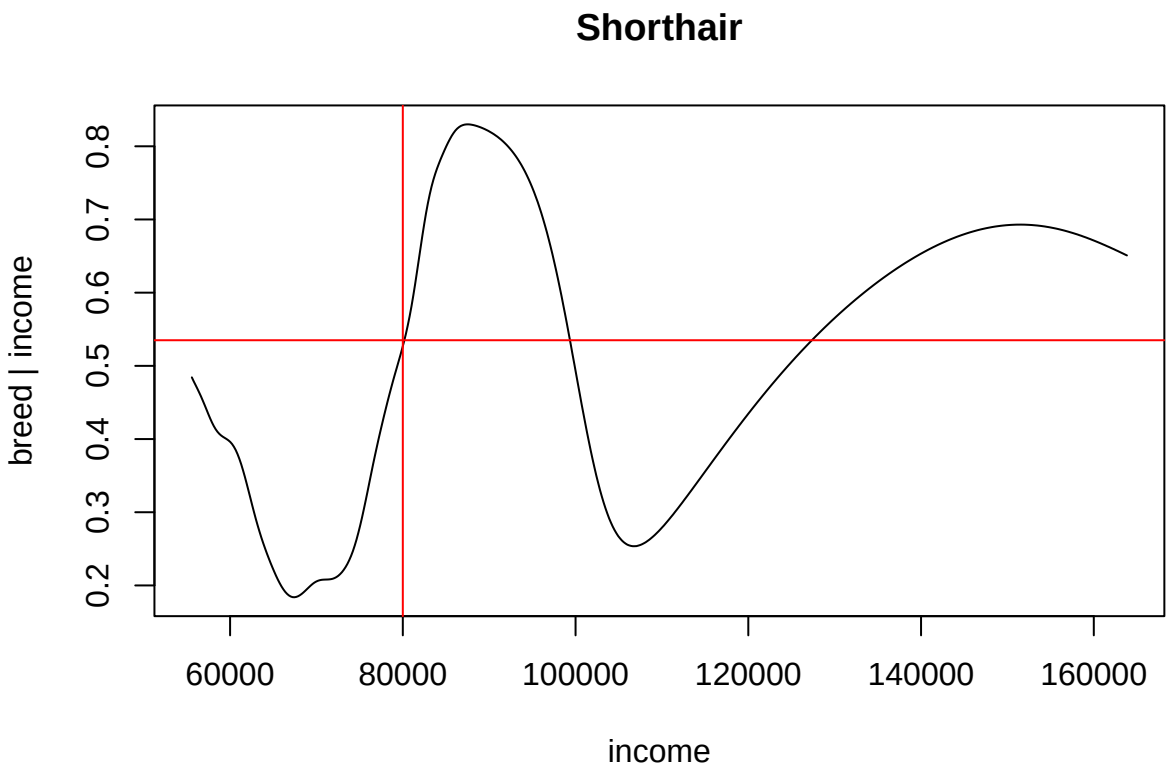
```
## [1] 0.1677019
```

This means that 45.96% of the cats in the data set are Bengals, 37.27% are Shorthairs, and 16.77% are Burmese.

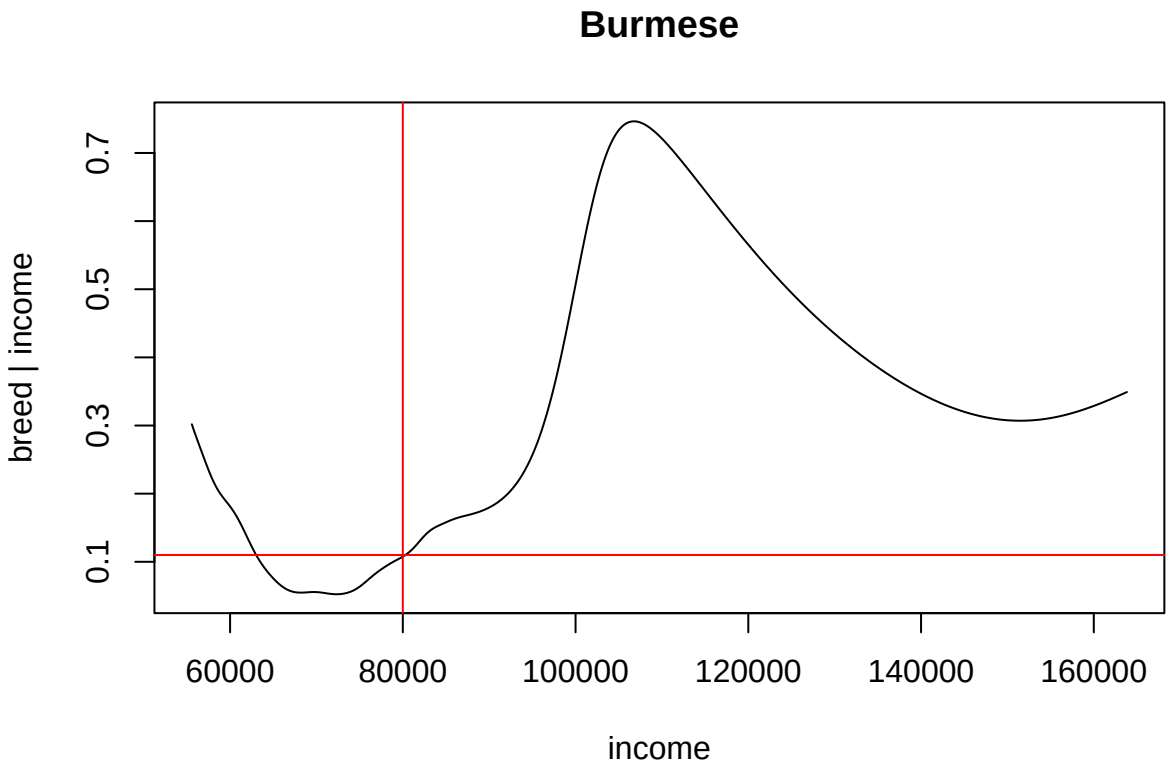
```
plot(f$x, f1$y*p1/(p1*f1$y+p2*f2$y+p3*f3$y), type = "l", main = paste('Bengal'), xlab = "income", ylab =
abline(v = 80000, col = "red")
abline(h = 0.36, col = "red")
```




```
plot(f$x, f2$y*p2/(p1*f1$y+p2*f2$y+p3*f3$y), type = "l", main = paste('Shorthair'), xlab = "income", yla
abline(v = 80000, col = "red")
abline(h = 0.535, col = "red")
```




```
plot(f$x, f3$y*p3/(p1*f1$y+p2*f2$y+p3*f3$y), type = "l", main = paste('Burmese'), xlab = "income", ylab
abline(v = 80000, col = "red")
abline(h = 0.11, col = "red")
```



Given that the family income is $80,000 from the graphs above we can make the following estimates:

Bengal: 0.36

Shorthair: 0.535

Burmese: 0.11

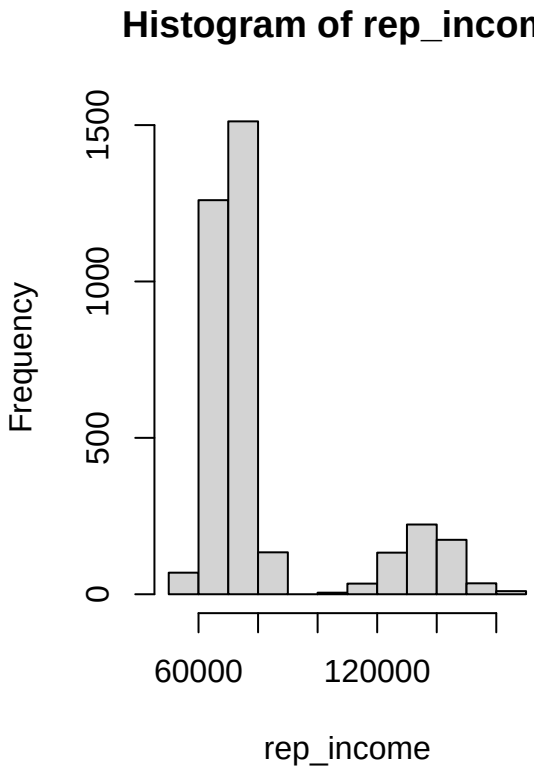
Gestation period and Income

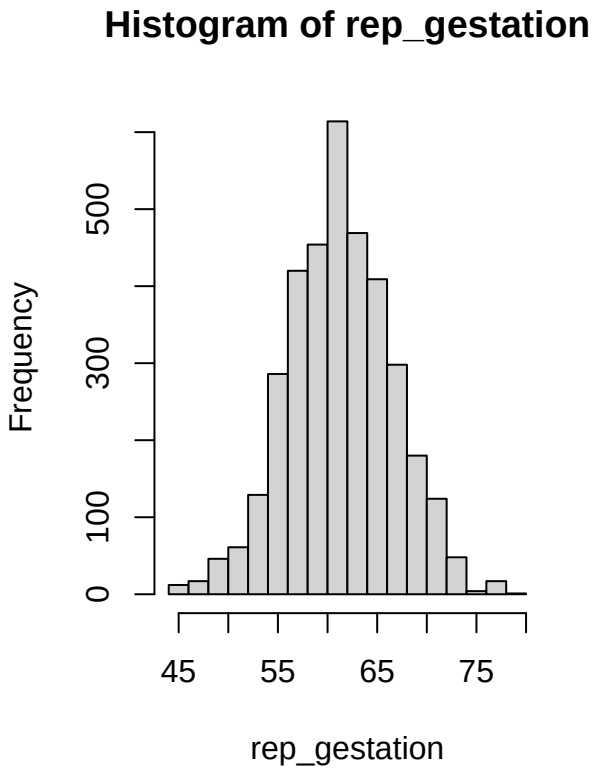
There is a belief that family income and the mental state of a cat are correlated, and that mental state is dependent on its time in the womb. Provide a model for the gestation period conditioned on the family income and estimate the coefficients to determine if there is a relationship. If any unobserved variables are identified, speculate what they might represent.

Since, in our dataset each row contains equal to or more than one cat birth, we need to replicate the rows for both income and gestation with respect to the litter size.

```
rep_gestation <- rep(data$gestation, data$litter_size)
rep_income <- rep(data$income, data$litter_size)
```

```
par(mfrow = c(1, 2))
hist(rep_income)
hist(rep_gestation)
```



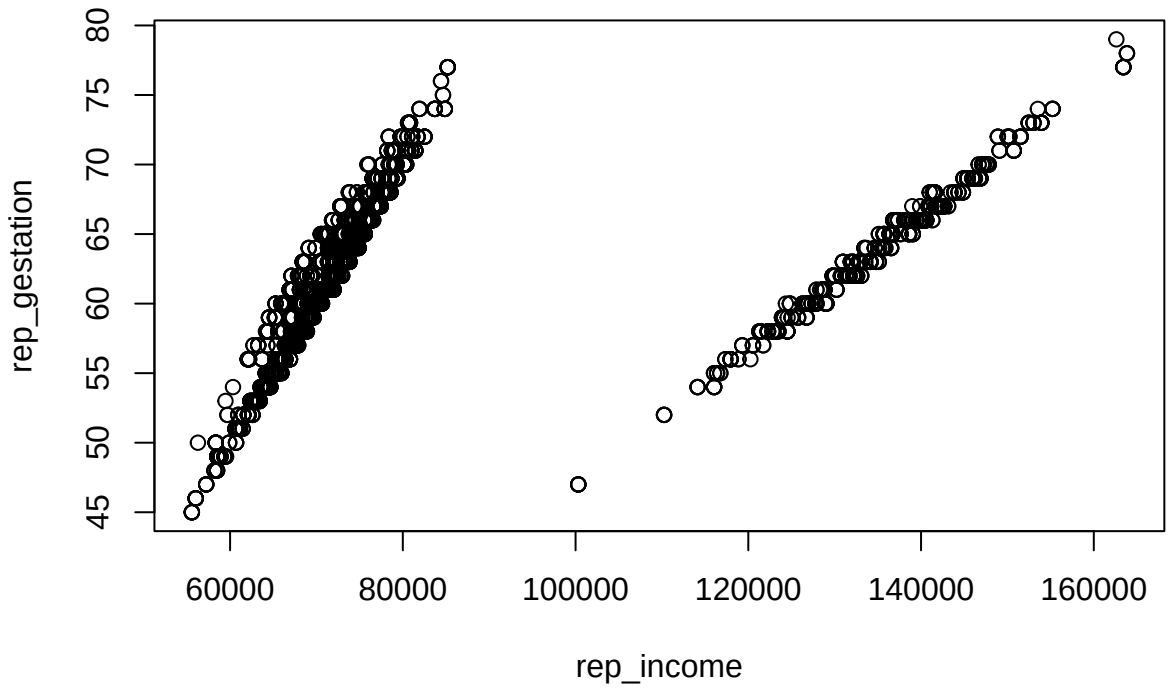


From the histogram of rep_income we can observe that most cats are born in families with an annual income of between $60,000 and $80,000.

The gestation histogram shows that most of the cats have a gestation period of 50 to 70 days.

For further investigation, we will produce a scatter plot to check if there is any relationship between family income and gestation period.

```
plot(rep_income, rep_gestation)
```



From the initial scatter plot above, we can observe that there is a positive relationship between the family income and gestation period but there seems to be two separate clusters. For further investigation we need to fit our data into a regression mixture model.

```
fit <- normalmixEM(rep_income, k = 2)
```

```
## number of iterations= 11
```

```
cov_matrices <- fit$Sigma

for (i in 1:2) {
  is_different_sigma <- !identical(cov_matrices[[i]], cov_matrices[[1]])
  cat("Component", i, "has a different sigma:", is_different_sigma, "\n")
}
```

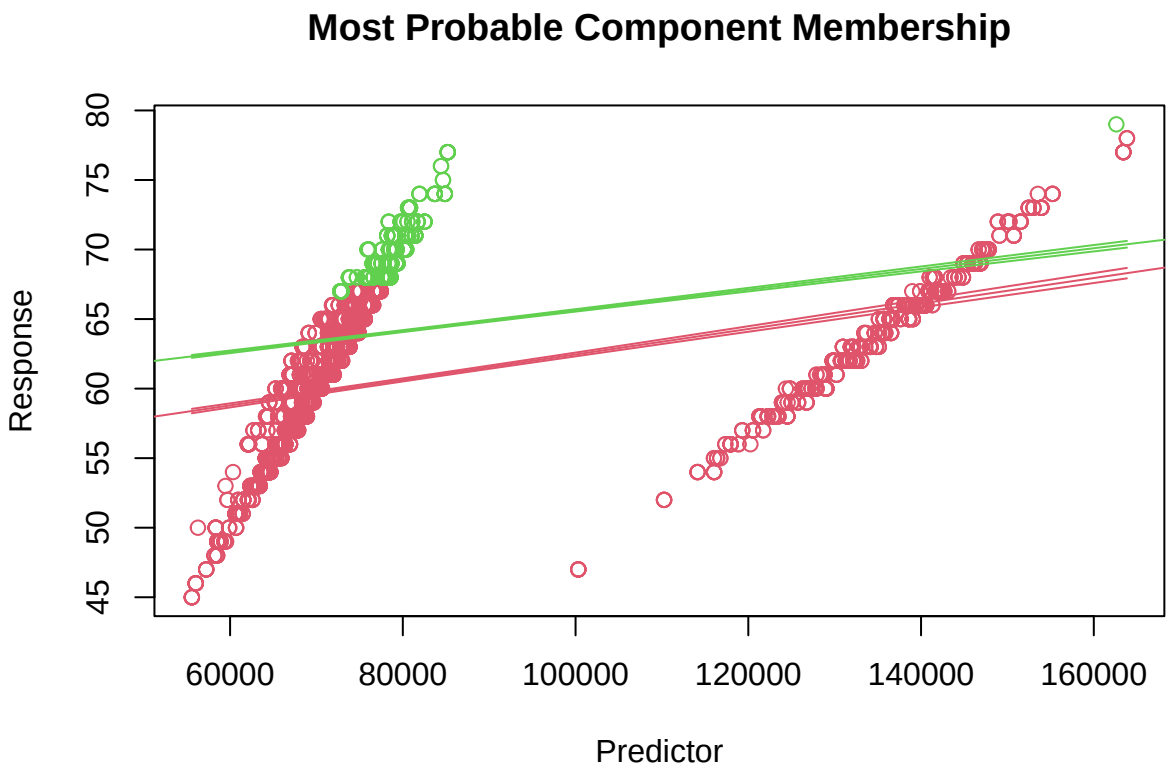
```
## Component 1 has a different sigma: FALSE
## Component 2 has a different sigma: FALSE
```

From the above, we can observe that rep_income have a single variance. Thus for the reg_model_1 where we have used regmixEM to fit our model, we have used arbvar = FALSE.

```
reg_model_1 <- regmixEM(x = rep_income, y = rep_gestation, k = 2, arbvar = FALSE)
```

```
## WARNING! NOT CONVERGENT!
## number of iterations= 10000
```

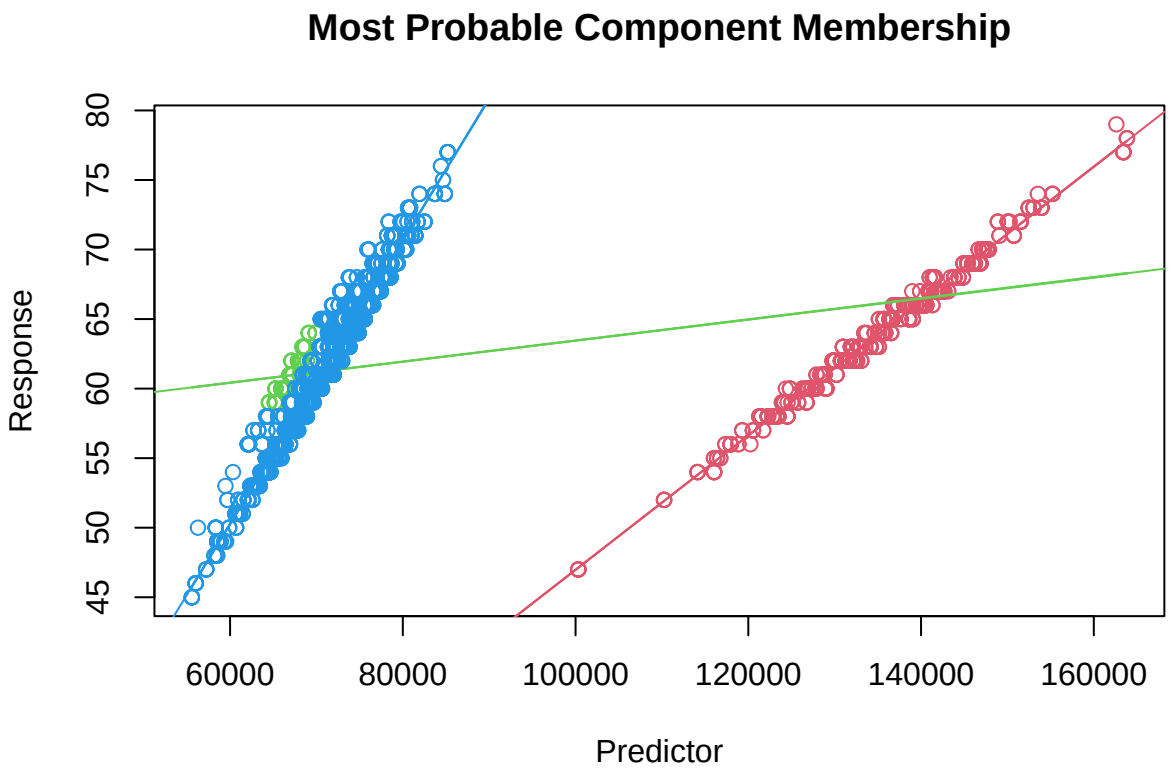
```
plot(reg_model_1, whichplots = 2)
```




```
reg_model_2 <- regmixEM(x = rep_income, y = rep_gestation, k = 3, arbvar = FALSE)
```

```
## number of iterations= 1093
```

```
plot(reg_model_2, whichplots = 2)
```




```
summary(reg_model_1)
```

```
## summary of regmixEM object:
##             comp 1      comp 2
## lambda 7.11668e-01 2.88332e-01
## sigma  4.51731e+00 4.51731e+00
## beta1  5.32982e+01 5.81781e+01
## beta2  9.15715e-05 7.44635e-05
## loglik at estimate:  -10712.03
```

```
summary(reg_model_2)
```

```
## summary of regmixEM object:
##              comp 1      comp 2       comp 3
## lambda  0.156604604 5.89944e-02   0.78440097
## sigma   0.997745884 9.97746e-01   0.99774588
## beta1  -1.325692163 5.58773e+01 -10.87526371
## beta2   0.000482996 7.57625e-05   0.00101913
## loglik at estimate:  -7022.691
```

```
aic <- c(-2*reg_model_1$loglik+2*(3*3-1),
         -2*reg_model_2$loglik+2*(3*4-1))
aic
```

```
## [1] 21440.05 14067.38
```

From the summary above, we can observe that the loglik at estimate for reg_model_1 is -1.0712026×10^4 and for reg_model_2 is -7022.6906062.

We have also calculated the AIC for the two models and we can see that the AIC for reg_model_2 is lower referring that it is a better model.

Coefficients:

```
## summary of regmixEM object:
##              comp 1      comp 2       comp 3
## lambda  0.156604604 5.89944e-02   0.78440097
## sigma   0.997745884 9.97746e-01   0.99774588
## beta1  -1.325692163 5.58773e+01 -10.87526371
## beta2   0.000482996 7.57625e-05   0.00101913
## loglik at estimate:  -7022.691
```

Lambda refers to the mixing proportions for each component.

Sigma refers to the standard deviation for each component.

Beta1 refers to the intercept and beta2 refers to the slope for each component.

In reg_model_2 comp 2 which has a very low lambda value is an unobserved variable. It could be due to one or more of the following issues:

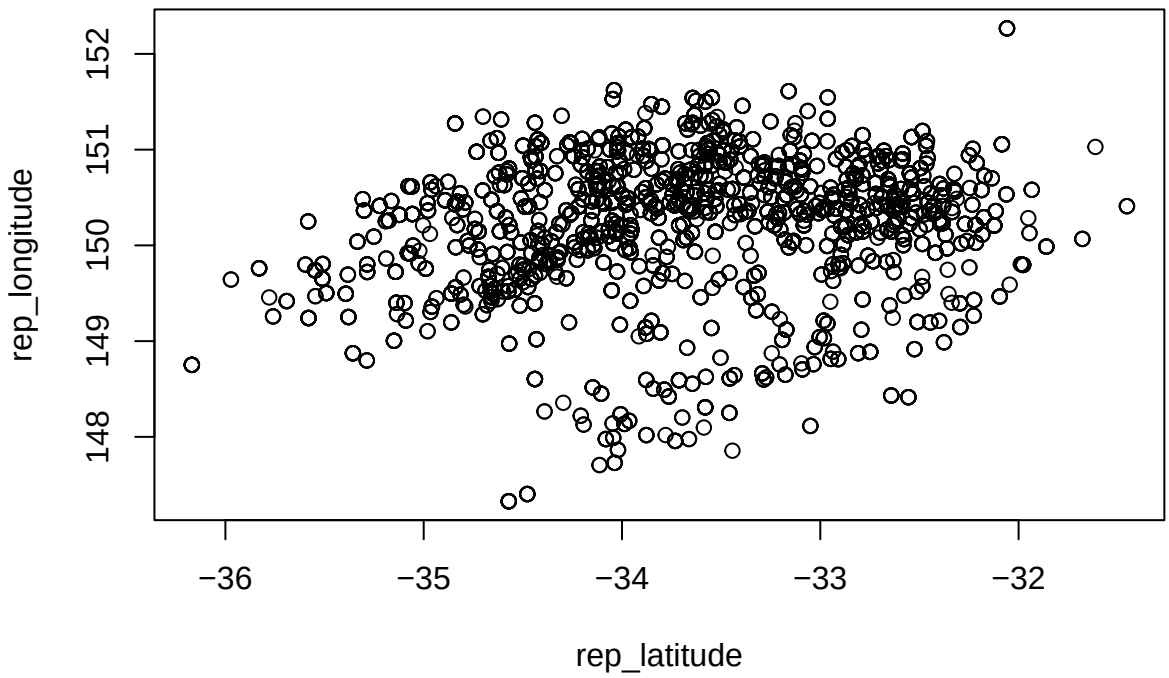
- health
- diet
- environment
- measurement error

Birth Suburbs

The RCSS want a simple model of birth place of the cats, to identify which suburbs produce the most cats. Provide a model of the density of the of the birth locations, and provide the fitted coefficients. Use the model to identify if more cats are likely to be born in Sydney or Parramatta.

```
rep_latitude <- rep(data$latitude, data$litter_size)
rep_longitude <- rep(data$longitude, data$litter_size)
rep_lat_long_df <- data.frame(rep_latitude, rep_longitude)
```

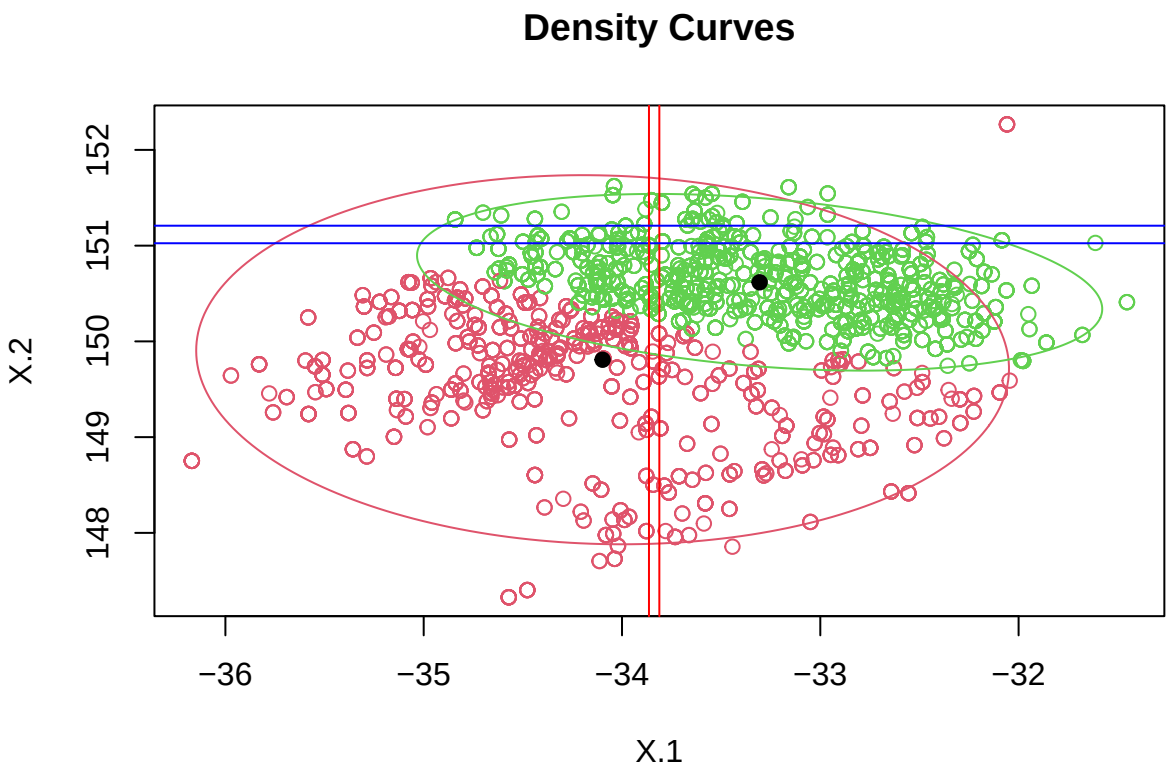
```
plot(rep_lat_long_df)
```




```
reg_model_3 <- mvnormalmixEM(rep_lat_long_df, k = 2)
```

```
## number of iterations= 149
```

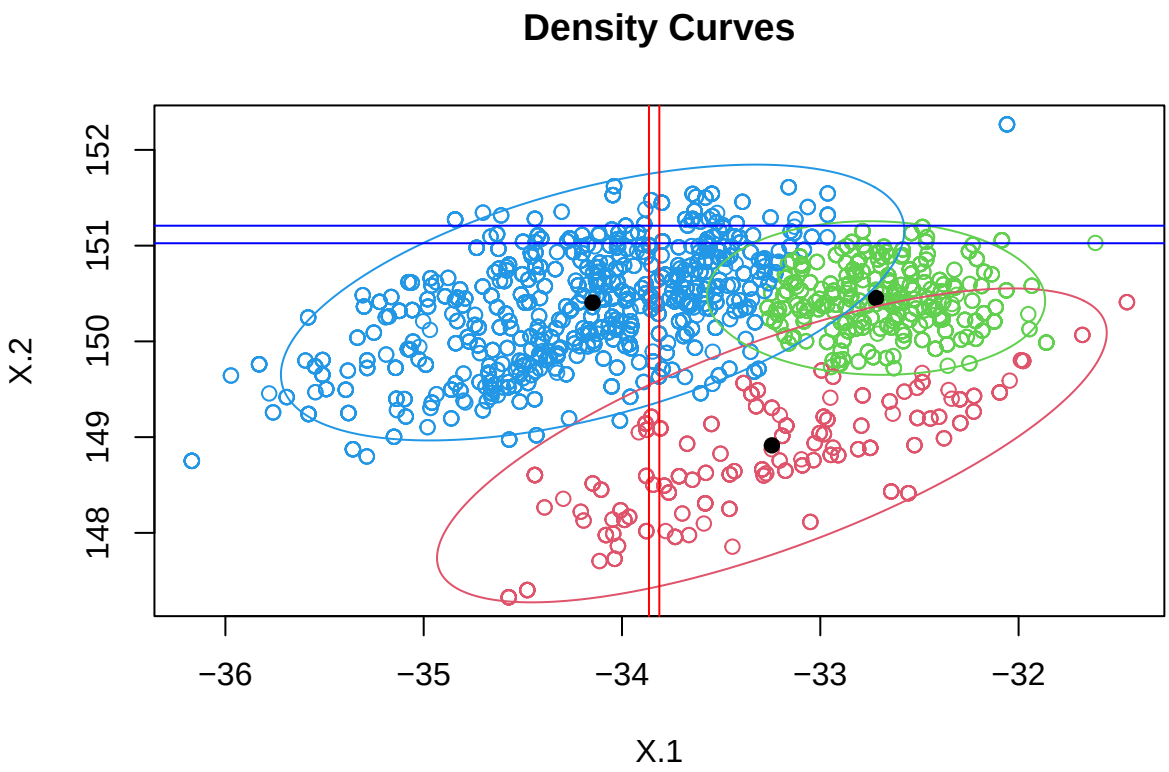
```
plot(reg_model_3, whichplots = 2)
abline(v = -33.81167161635268, col = "red")
abline(h = 151.02512573859647, col = "blue")
abline(v = -33.86395922654538, col = "red")
abline(h = 151.20818397512764, col = "blue")
```




```
reg_model_4 <- mvnormalmixEM(rep_lat_long_df, k = 3)
```

```
## number of iterations= 255
```

```
plot(reg_model_4, whichplots = 2)
abline(v = -33.81167161635268, col = "red")
abline(h = 151.02512573859647, col = "blue")
abline(v = -33.86395922654538, col = "red")
abline(h = 151.20818397512764, col = "blue")
```




```
summary(reg_model_3)
```

```
## summary of mvnormalmixEM object:
##            comp 1     comp 2
## lambda   0.453475   0.546525
## mu1    -34.097865 -33.305925
## mu2    149.808418 150.617335
## loglik at estimate:  -8042.824
```

```
summary(reg_model_4)
```

```
## summary of mvnormalmixEM object:
##            comp 1   comp 2     comp 3
## lambda   0.112552   0.2668   0.620648
## mu1    -33.244099 -32.7188 -34.148176
## mu2    148.913808 150.4530 150.405880
## loglik at estimate:  -7696.921
```

From the summary above, we can observe that the loglik at estimate for reg_model_1 is -8042.8242515 and for reg_model_2 is -7696.9214794. Since the loglik at estimate for reg_model_4 is higher, we can suggest that it is a better model.

Since the coordinates for Parramatta intersects closer to the center of the cluster, we can suggest that cats are likely to be born in Parramatta.