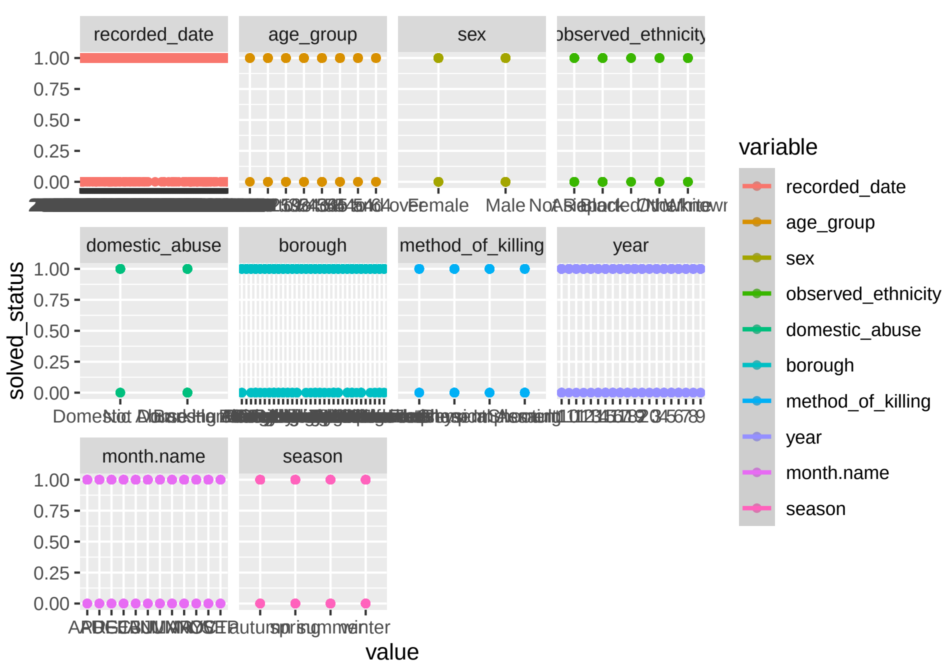
Exploratory Data Analysis

```
homicides = read.csv('homicides.csv')
data = read.csv('homicides.csv')
data$solved_status = as.integer(data$solved_status == "Solved")

## 75% of the sample size
smp_size <- floor(0.75 * nrow(data))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(data)), size = smp_size)

train <- data[train_ind, ]
test <- data[-train_ind, ]
```

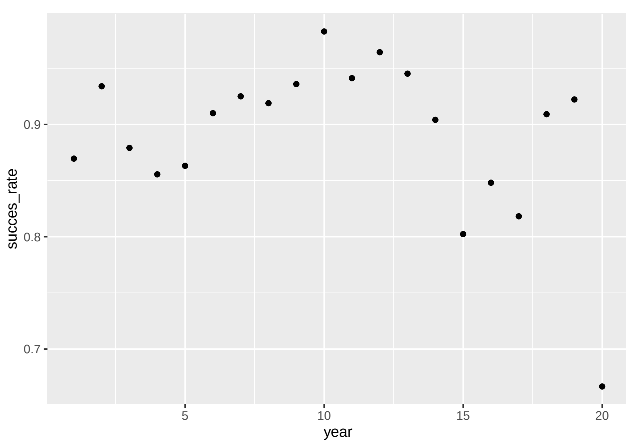
First, I separate the data set into a training and test set, in a 75-25 split. I have also randomly shuffled the samples.

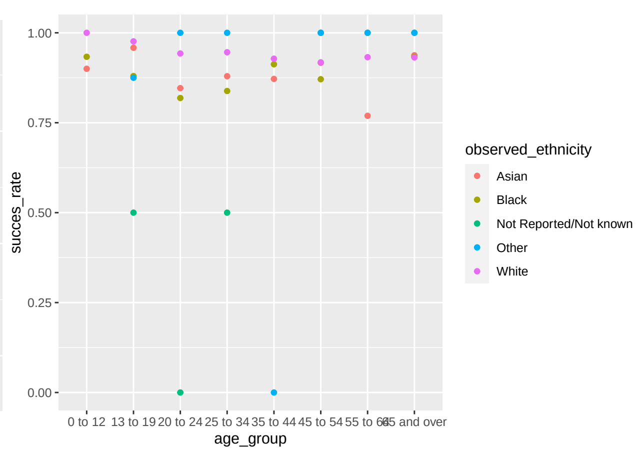
```
# Plotting Data
murderData = melt(train, id.vars = 'solved_status')
ggplot(murderData) +
  geom_point(aes(x=value, y= solved_status , colour=variable)) +
  geom_smooth(aes(x=value, y= solved_status, colour=variable),method = lm) +
  facet_wrap(~variable, scales="free_x")

## `geom_smooth()` using formula = 'y ~ x'
```

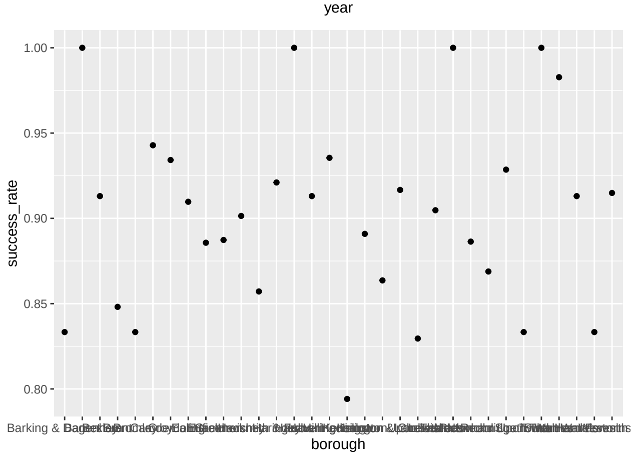


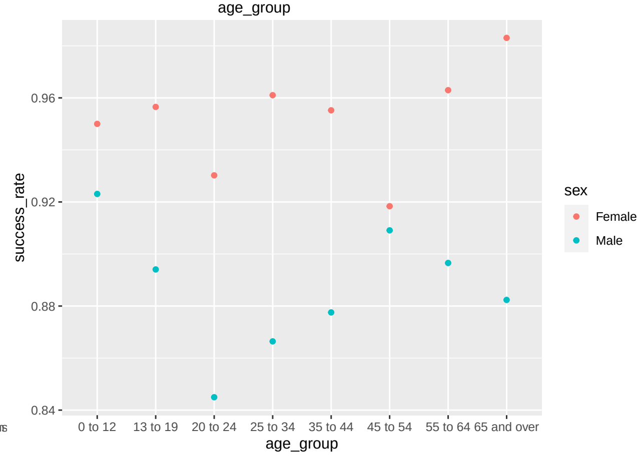
Looking at this plot does not tell us much as the output is binary. As such, we will look at more specific plots to see any correlation in our data.

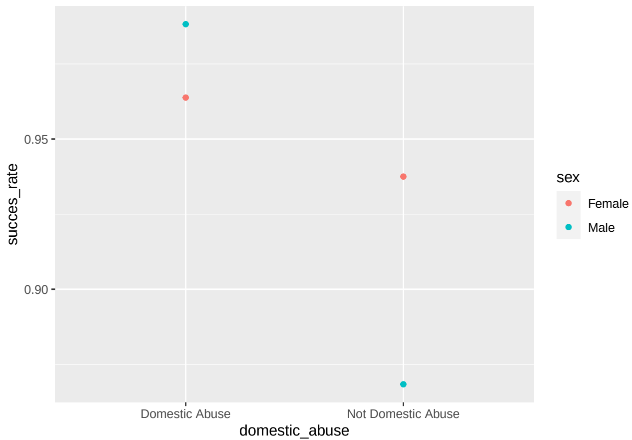
```r
data_by_ethn = train %>% group_by(observed_ethnicity, age_group) %>% summarize
    (succes_rate = mean(solved_status))
data_by_kill = train %>% group_by(method_of_killing, age_group) %>% summarize(
    success_rate = mean(solved_status))
data_by_bor = train %>% group_by(borough) %>% summarize(success_rate = mean(
    solved_status))
data_by_sex = train %>% group_by(age_group, sex) %>% summarize(success_rate =
    mean(solved_status))
data_by_da = train %>% group_by(domestic_abuse, sex) %>% summarize(succes_rate
    = mean(solved_status))
data_by_seas = train %>% group_by(year) %>% summarize(succes_rate = mean(
    solved_status))
ggplot(data_by_seas, aes(x = year, y = succes_rate)) + geom_point()
ggplot(data_by_ethn, aes(x = age_group, y = succes_rate, color = observed_
    ethnicity)) + geom_point()
ggplot(data_by_bor, aes(x = borough, y = success_rate, )) + geom_point()
ggplot(data_by_sex, aes(x = age_group, y = success_rate, color = sex)) + geom_
    point()
ggplot(data_by_da, aes(x = domestic_abuse, y = succes_rate, color = sex)) +
    geom_point()
ggplot(data_by_kill, aes(x = age_group, y = success_rate, color = method_of_
    killing)) + geom_point()
```

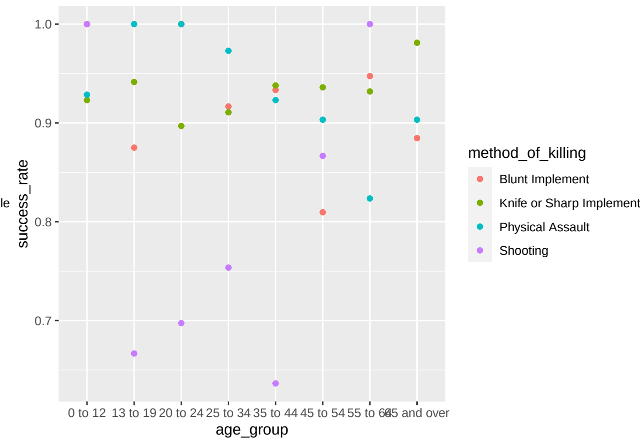












Looking at these plots, it appears that year doesn't have much of an effect on murders getting solved, similarly for boroughs. However, murders seem less likely to be solved given the individuals ethnicity and age group, particularly 13 to 19 to 35 to 44. Sex also seems to have an influence with males of a similar age group seeming less likely to have murder cases solved. Again, males not in a domestic abuse related crime also seem less likely to have their murder case solved. Finally, method of killing, particularly shootings appear to have less cases solved, again in a similar age group. We will now use our findings to build our Bayesian Model.

Bayesian Model

Given our exploratory data analysis, the variables I have chosen to regress on are observed_ethnicity, sex, domestics_abuse and method_of_killing. I have decided to use age_group as my grouping variable. More formally, the model is :

$$y_i = \begin{cases} 1 & \text{if Solved} \\ 0 & \text{if Unsolved} \end{cases}$$

$$x_{ij1} := \text{observed ethnicity}$$

$$x_{ij2} := \text{sex}$$

$$x_{ij3} := \text{domestic abuse}$$

$$x_{ij4} := \text{method of killing}$$

Let π_{ij} be the probability that a murder case, i in age group j be solved i.e $y_{ij} = 1$

$$y_{ij}|\beta j, x_{ij} \sim Bern(\pi_{ij})$$

$$\text{with } log(\frac{\pi_{ij}}{1-\pi_{ij}}) = b_0 + b_1 x_{ij1} + b_2 x_{ij2} + b_3 x_{ij3} + b_4 x_{ij4} + \beta_{0j} + \beta_{1j} x_{ij1} + \beta_{2j} x_{ij2} + \beta_{3j} x_{ij3} + \beta_{4j} x_{ij4}$$

$$\beta j \sim N(0, \Sigma)\ ,\ \pi(b, \Sigma)$$

I will now look to set my priors and explain my reasoning. As the predictors are categorical variables, I must set a prior for each categorical value.

$$b_0 \sim N(2, 4)$$

$$b_{1,white} \sim N(0, 0.6),\ b_{1,black} \sim N(0, 0.4),\ b_{1,Other} \sim N(0, 1),\ b_{1,Notknown} \sim N(0, 1)$$

$$b_{2,male} \sim N(0, 0.2)$$

$$b_{3,notDomesticAbuse} \sim N(0, 0.1)$$

$$b_{4,knife} \sim N(0, 0.3),\ b_{4,phys} \sim N(0, 0.3),\ b_{4,shoot} \sim N(0, 1)$$

$$\sigma_0 \sim N(0, 0.1)$$

$$\sigma_{1,white} \sim N(0, 0.05)\ \sigma_{1,black} \sim N(0, 0.1)\ \sigma_{1,Other} \sim N(0, 1)\ \sigma_{1,Notknown} \sim N(0, 0.1)$$

$$\sigma_{2,male} \sim N(0, 0.05)$$

$$\sigma_{3,notDomestiAbuse} \sim N(0, 0.1)$$

$$\sigma_{4,knife} \sim N(0, 0.1)\ \sigma_{4,phys} \sim N(0, 0.1)\ \sigma_{4,shoot} \sim N(0, 0.5)$$

For all priors, I have set mean 0 in the event that they do not have an impact on our baseline. For this model, looking at the data, I have assumed the average murder solving rate to be about 88%, but that can range from 0.01-99.99%. As such, I have set my intercept prior to be N(2,4) (logit(0.88) = 2). Looking at our other priors, these coefficients increase or decrease our baseline. For the observed ethnicity, I have assumed that white may increase our baseline to about 95%, as such, b1_white = logit(0.88) + logit(0.07) and |b1_white| is about 0.6. Similarly, I assume black affects the baseline, maybe decreasing it at most to 76%. As such, I have set b2 to N(0,0.4). I do however believe that Other and Not Known have a strong impact on the baseline, looking at extreme level so two standard deviations, I think It could reduce the probability of the murder being solved to 50%, as such, 2*b1_Other = logit(0.88) - logit(0.38) and so |b1_other| is approximately 1. I don't believe, domestic abuse has much impact, if any, and so that may be worth looking at during our sensitivity analysis. For the method of killings, I believe that different methods make the baseline vary between 80-99% and so have set the priors accordingly. I believe shooting heavily decreases the baseline, and so again, have set a large prior. For the sigmas, I think the age group will have an effect on the baseline. I believe certain age groups, for example 18-30 year old have less probability of having their case solved, whilst other age groups, like 0-12 and 40+ are less likely to be involved in crime and so have increased probability of having their case solved. Similarly, I think shooting and knife crime will be more influential for certain age groups more likely to be involved in crime. I'm not sure about the impact of certain age groups on black/white ethnicity so have set low priors. I do think the age group could have a higher impact given the age group, as older people are more likely to be in relationships, although lower age groups are more likely to suffer domestic abuse from family.

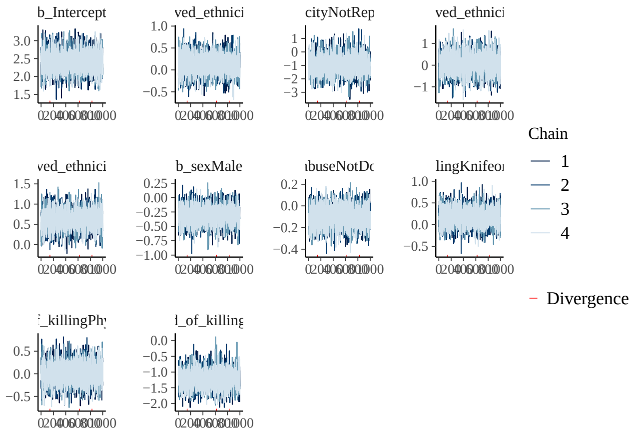
```
prior_int = set_prior("normal(2, 4)", class="Intercept")
prior_b = set_prior("normal(0,0.1)", class = "b")
prior_ethn_black = set_prior("normal(0,0.4)", class="b", coef="observed_
    ethnicityBlack")
prior_ethn_white = set_prior("normal(0,0.6)", class="b", coef="observed_
    ethnicityWhite")
prior_ethn_nk = set_prior("normal(0,1)", class="b", coef="observed_
    ethnicityNotReportedDNotknown")
prior_ethn_other = set_prior("normal(0,0.5)", class="b", coef="observed_
    ethnicityOther")
prior_sexM = set_prior("normal(0,0.2)", class = "b", coef="sexMale")
prior_daN = set_prior("normal(0,0.1)", class = "b", coef="domestic_
    abuseNotDomesticAbuse")
prior_kill_sharp = set_prior("normal(0,0.3)", class = "b", coef = "method_of_
    killingKnifeorSharpImplement")
prior_kill_phys = set_prior("normal(0,0.3)", class = "b", coef = "method_of_
    killingPhysicalAssault")
prior_kill_shoot = set_prior("normal(0,1)", class = "b", coef = "method_of_
    killingShooting")

sd_priorsI = set_prior("normal(0,0.1)", class = "sd",group = "age_group", coef
     = "Intercept")
sd_priors_kill = set_prior("normal(0,0.1)", class = "sd", group = "age_group",
     coef = "method_of_killingKnifeorSharpImplement")
sd_priors_Phys = set_prior("normal(0,0.1)", class = "sd", group = "age_group",
    coef = "method_of_killingPhysicalAssault")
```

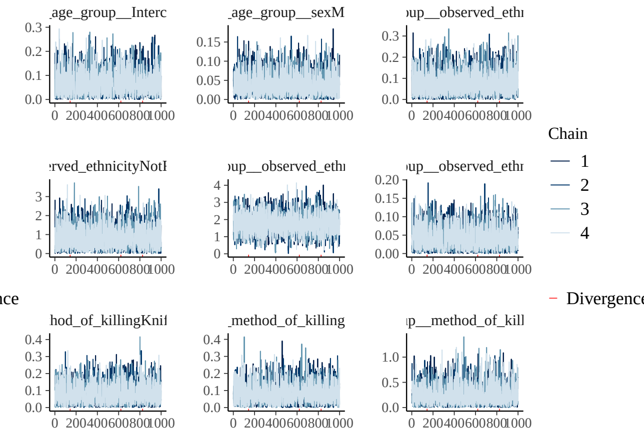
```
sd_priors_Shot = set_prior("normal(0,0.5)", class = "sd", group = "age_group",
    coef = "method_of_killingShooting")
sd_priors_B = set_prior("normal(0,0.1)", class = "sd", group = "age_group",
    coef = "observed_ethnicityBlack")
sd_priors_Nr = set_prior("normal(0,1)", class = "sd",group = "age_group", coef
     = "observed_ethnicityNotReportedDNotknown")
sd_priors_Oth = set_prior("normal(0,1)", class = "sd",group = "age_group",
    coef = "observed_ethnicityOther")
sd_priors_W = set_prior("normal(0,0.05)", class = "sd", group = "age_group",
    coef = "observed_ethnicityWhite")
sd_priors_M = set_prior("normal(0,0.05)", class = "sd",group = "age_group",
    coef = "sexMale")

priors = c(prior_int, prior_ethn_black, prior_ethn_white, prior_daN, prior_
    sexM, prior_kill_phys, prior_kill_sharp
            ,prior_kill_shoot, prior_b, sd_priorsI, sd_priors_kill, sd_priors_
               Phys, sd_priors_Shot, sd_priors_B,
            sd_priors_Nr, sd_priors_Oth, sd_priors_W, sd_priors_M, prior_ethn_
               nk, prior_ethn_other)

murder_fit = brm(solved_status ~ observed_ethnicity + sex  + domestic_abuse +
    method_of_killing + (sex+observed_ethnicity+method_of_killing|age_group),
                 family = bernoulli(link = 'logit'), prior = priors, data =
                    train)

mcmc_plot(murder_fit, type = "trace", variable = "^b_", regex = TRUE)
mcmc_plot(murder_fit, type = "trace", variable = "^sd_", regex = TRUE)
summary(murder_fit)$fixed

##                                           Estimate  Est.Error   l-95% CI
## Intercept                               2.42709699 0.28385273  1.8812361
## observed_ethnicityBlack                 0.13785358 0.21740278 -0.2927370
## observed_ethnicityNotReportedDNotknown -0.94220544 0.68031029 -2.2101223
## observed_ethnicityOther                 0.05636434 0.45787465 -0.8180959
## observed_ethnicityWhite                 0.61990719 0.23406076  0.1656094
## sexMale                                -0.30660497 0.15608844 -0.6189643
## domestic_abuseNotDomesticAbuse         -0.09907436 0.09316992 -0.2865779
## method_of_killingKnifeorSharpImplement  0.18772884 0.20715423 -0.2224322
## method_of_killingPhysicalAssault        0.01592671 0.22952304 -0.4252889
## method_of_killingShooting              -1.21366675 0.29501984 -1.7722593
##                                           u-95% CI      Rhat Bulk_ESS
    Tail_ESS
## Intercept                               2.99518707 1.0009286 5748.023
    2628.718
## observed_ethnicityBlack                 0.55874150 1.0023516 6175.956
    3415.555
## observed_ethnicityNotReportedDNotknown  0.45369267 1.0007107 5427.226
    3217.037
## observed_ethnicityOther                 0.96970349 1.0015842 6670.311
    3193.302
## observed_ethnicityWhite                 1.07918997 1.0011598 6031.652
    3144.784
## sexMale                                -0.00571536 1.0022685 6420.875
    2383.445
```

```
## domestic_abuseNotDomesticAbuse            0.07715082  1.0008377  6924.272
   2851.025
## method_of_killingKnifeorSharpImplement   0.58219050  0.9997321  5646.613
   3049.261
## method_of_killingPhysicalAssault          0.48552541  1.0004607  6454.934
   2352.246
## method_of_killingShooting                -0.62347977  1.0009851  3856.945
   2133.285
```

```r
summary(murder_fit)$random$age_group[1:9,]
```

```
##                                                Estimate   Est.Error      l-95%
   CI
## sd(Intercept)                                0.06423022 0.04911462
   0.002097078
## sd(sexMale)                                  0.03891898 0.02908742
   0.001615251
## sd(observed_ethnicityBlack)                  0.07395016 0.05540327
   0.002825508
## sd(observed_ethnicityNotReportedDNotknown)   0.79776475 0.58137648
   0.033410041
## sd(observed_ethnicityOther)                  1.70987113 0.59379112
   0.644209184
## sd(observed_ethnicityWhite)                  0.03938879 0.03032878
   0.001560123
## sd(method_of_killingKnifeorSharpImplement)   0.07872195 0.05911087
   0.003411415
## sd(method_of_killingPhysicalAssault)         0.07625865 0.05924732
   0.003103907
## sd(method_of_killingShooting)                0.24715124 0.20033932
   0.008790673
##                                               u-95% CI      Rhat  Bulk_ESS
## sd(Intercept)                                0.1834930 1.0015303 3077.239
## sd(sexMale)                                  0.1079448 1.0009882 3540.814
## sd(observed_ethnicityBlack)                  0.2083300 1.0008871 3718.534
## sd(observed_ethnicityNotReportedDNotknown)   2.1676393 0.9999626 2960.350
## sd(observed_ethnicityOther)                  2.9829758 0.9996624 3718.907
## sd(observed_ethnicityWhite)                  0.1117189 1.0011129 3348.729
## sd(method_of_killingKnifeorSharpImplement)   0.2206792 1.0001061 3032.878
## sd(method_of_killingPhysicalAssault)         0.2181387 1.0002986 2653.871
## sd(method_of_killingShooting)                0.7525961 0.9996971 2661.531
##                                              Tail_ESS
## sd(Intercept)                                1798.399
## sd(sexMale)                                  2247.053
## sd(observed_ethnicityBlack)                  2447.269
## sd(observed_ethnicityNotReportedDNotknown)   2232.640
## sd(observed_ethnicityOther)                  1872.772
## sd(observed_ethnicityWhite)                  1956.144
## sd(method_of_killingKnifeorSharpImplement)   2275.411
## sd(method_of_killingPhysicalAssault)         1994.770
## sd(method_of_killingShooting)                2654.235
```





Looking at the traceplots, everything seems well mixed. The Rhat values are also fine and very close to 1. I am happy that the model has converged. Looking briefly at our summary, at the population level, our intercept is at 2.4 which is about 91%, ethnicity unknown reduces considerably our baseline, as well as method of killing shooting. Method of Killing knife however seems to increase our baseline probability. Observed ethnicity white appears to increase the baseline. Sex male appears to decrease our baseline probability. Looking at the grouping, ethnicity other and not reported/unknown seem to be heavily impacted by the age group. Method of killing shooting also appears to be impacted by the age group.

Here we will look at how well our model is at predicting solved/unsolved murder cases. To assess the model, we will look at a confusion matrix, which will tell us how often it is right, and which type of errors the model is making.

```
preds <- predict(murder_fit , newdata=test)
head(preds)
```

```
##        Estimate Est.Error Q2.5 Q97.5
## [1,]    0.95200 0.2137930    0     1
## [2,]    0.89950 0.3007031    0     1
## [3,]    0.89975 0.3003706    0     1
## [4,]    0.96000 0.1959837    0     1
## [5,]    0.95750 0.2017521    0     1
## [6,]    0.71500 0.4514709    0     1
```

```
a_classifier <- preds[,"Estimate"]>0.909
ConfusionMatrix <- function(Classifier, Truth){
  if(!(length(Classifier)==length(Truth)))
    stop("Make the length of your vector of predictions the same as the length
          of the truth")
  if(is.logical(Classifier))
    Classifier <- as.integer(Classifier)
  WhichClass0s <- which(Classifier < 1)
  ZeroCompare <- Truth[WhichClass0s]
  Predicted0 <- c(length(ZeroCompare)-sum(ZeroCompare), sum(ZeroCompare))
  WhichClass1s <- which(Classifier >0)
  OnesCompare <- Truth[WhichClass1s]
  Predicted1 <- c(length(OnesCompare)-sum(OnesCompare), sum(OnesCompare))
  ConMatrix <- cbind(Predicted0,Predicted1)
  row.names(ConMatrix) <- c("Actual 0", "Actual 1")
  colnames(ConMatrix) <- c("Pred 0", "Pred 1")
  ConMatrix
}
```

```r
conmat <- ConfusionMatrix(a_classifier , test$solved_status)
conmat
```

```
##           Pred 0 Pred 1
## Actual 0      31     26
## Actual 1     149    327
```

```r
sum(diag(conmat))/sum(conmat)
```

```
## [1] 0.6716698
```

First we look at our predictions, we see that the model returns a Monte Carlo estimate. This is the probability that the predicated value will be 1 (or Solved). We set everything with posterior predictive probability >0.91 to be 1 and 0 otherwise. We see that our model has approximately 68% accuracy. Looking closer, we see that the model has 54% accuracy for unsolved murders, and 70% accuracy for solved murders. Although the accuracy isn't very high, I am happy this model provides an adequate fit to the data.

Sensitivity

Here we will look at what our model is sensitive to. I will analyze the impact of removing method of killing from our model.

```r
priors = c(prior_int , prior_ethn_black , prior_ethn_white , prior_daN, prior_
    sexM, prior_b, sd_priorsI , sd_priors_B,
             sd_priors_Nr, sd_priors_Oth, sd_priors_W, sd_priors_M, prior_ethn_
                 nk, prior_ethn_other)

sens_fit = brm(solved_status ~ observed_ethnicity + sex  + domestic_abuse + (
    sex+observed_ethnicity|age_group),
                      family = bernoulli(link = 'logit'), prior = priors, data =
                          train)

mcmc_plot(sens_fit , type = "trace", variable = "^b_", regex = TRUE)
```

```
## No divergences to plot.
```

```r
mcmc_plot(sens_fit , type = "trace", variable = "^sd_", regex = TRUE)
```

```
## No divergences to plot.
```

```r
summary(sens_fit)$fixed
```

```
##                                          Estimate  Est.Error   l-95% CI
## Intercept                              2.43401445 0.22391510  2.0133772
## observed_ethnicityBlack               -0.11539033 0.20234124 -0.5140255
## observed_ethnicityNotReportedDNotknown -0.96810068 0.69529252 -2.2444016
## observed_ethnicityOther                0.03375651 0.45769280 -0.8667341
## observed_ethnicityWhite                0.61402678 0.22071928  0.1952449
## sexMale                               -0.36561854 0.15403138 -0.6676159
## domestic_abuseNotDomesticAbuse        -0.13496866 0.09355181 -0.3174232
##                                          u-95% CI      Rhat Bulk_ESS
##     Tail_ESS
## Intercept                              2.88406433 1.0023454 3734.204
##     2644.936
```

```
## observed_ethnicityBlack                    0.28093711  1.0002763  3228.277
   2808.909
## observed_ethnicityNotReportedDNotknown    0.50001924  0.9998171  3441.291
   2465.263
## observed_ethnicityOther                    0.91503951  1.0010110  3831.361
   2752.913
## observed_ethnicityWhite                    1.04566442  1.0024236  3422.422
   3291.358
## sexMale                                   -0.07447470  1.0006729  4572.985
   3161.157
## domestic_abuseNotDomesticAbuse             0.04271414  0.9996996  4447.610
   3056.032
```

```
summary(sens_fit)$random$age_group[1:9,]
```

```
##                                              Estimate   Est.Error        l
   -95% CI
## sd(Intercept)                               0.069745516 0.05198092
   0.003651357
## sd(sexMale)                                 0.040026565 0.02925499
   0.001672829
## sd(observed_ethnicityBlack)                 0.082366271 0.06025014
   0.003555059
## sd(observed_ethnicityNotReportedDNotknown)  0.796003658 0.56833929
   0.044265297
## sd(observed_ethnicityOther)                 1.759236551 0.59568593
   0.695555478
## sd(observed_ethnicityWhite)                 0.039171652 0.02847860
   0.001869774
## cor(Intercept,sexMale)                     -0.006610031 0.37589091
   -0.700571862
## cor(Intercept,observed_ethnicityBlack)     -0.003048651 0.36830340
   -0.690999395
## cor(sexMale,observed_ethnicityBlack)        0.004068667 0.37762436
   -0.708460543
##                                            u-95% CI      Rhat  Bulk_ESS
## sd(Intercept)                              0.1963118 1.0010981 2796.916
## sd(sexMale)                                0.1112084 1.0005447 2833.498
## sd(observed_ethnicityBlack)                0.2249160 1.0004386 2304.346
## sd(observed_ethnicityNotReportedDNotknown) 2.1582313 0.9995103 2333.817
## sd(observed_ethnicityOther)                3.0366843 1.0008260 2305.396
## sd(observed_ethnicityWhite)                0.1067742 0.9992229 3750.579
## cor(Intercept,sexMale)                     0.7016541 1.0023917 4611.079
## cor(Intercept,observed_ethnicityBlack)     0.6906803 1.0003031 4120.578
## cor(sexMale,observed_ethnicityBlack)       0.7008209 1.0011033 3784.272
##                                            Tail_ESS
## sd(Intercept)                              2678.637
## sd(sexMale)                                1878.752
## sd(observed_ethnicityBlack)                1526.245
## sd(observed_ethnicityNotReportedDNotknown) 2072.373
## sd(observed_ethnicityOther)                1127.760
## sd(observed_ethnicityWhite)                1853.514
## cor(Intercept,sexMale)                     2761.696
## cor(Intercept,observed_ethnicityBlack)     2768.303
```

```
## cor(sexMale,observed_ethnicityBlack)          2768.436
```

```
preds2 <- predict(sens_fit, newdata=test)
head(preds2)
```

```
##        Estimate Est.Error Q2.5 Q97.5
## [1,]   0.95300 0.2116653    0     1
## [2,]   0.86725 0.3393468    0     1
## [3,]   0.86625 0.3404261    0     1
## [4,]   0.94000 0.2375165    0     1
## [5,]   0.94675 0.2245597    0     1
## [6,]   0.87525 0.3304765    0     1
```

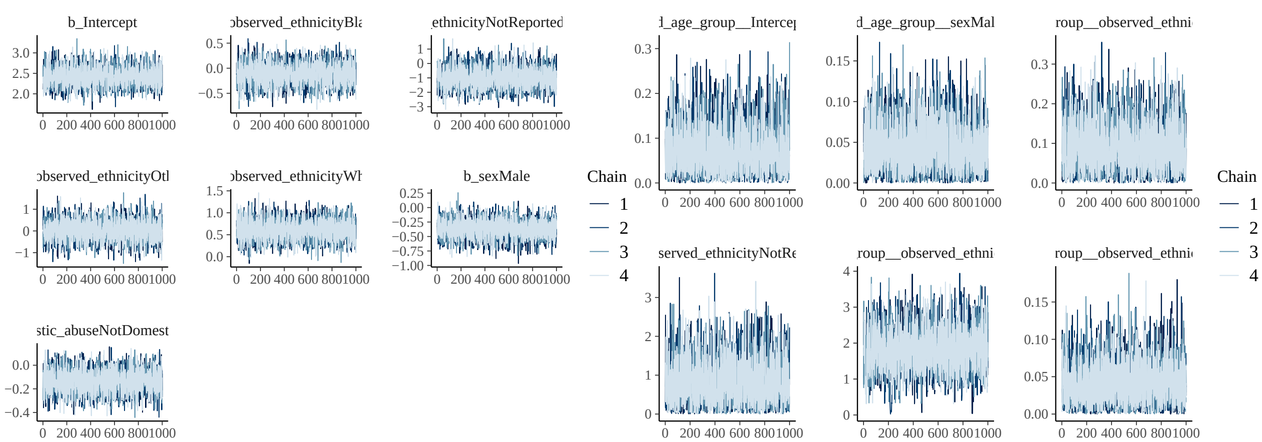
```
a_classifier <- preds2[,"Estimate"]>0.909
```

```
conmat <- ConfusionMatrix(a_classifier, test$solved_status)
conmat
```

```
##            Pred 0 Pred 1
## Actual 0       37     20
## Actual 1      225    251
```

```
sum(diag(conmat))/sum(conmat)
```

```
## [1] 0.5403377
```



Briefly looking at our sensitivity model, we see that the model has still converged. The intercept remains 2.4. Observed ethnicity black now appears to decrease our baseline and domestic abuse seems to have more impact. Our other coefficients appear unchanged. We also see that the predictive accuracy is lower at 54%. Although the predictive accuracy of unsolved murders is higher, the predicitive accuracy of solved murders has heavily decreased to 52%.

Inference

In this part, we will use the model to infer how the features of any particular homicide in London affect the probability that the case has been solved.

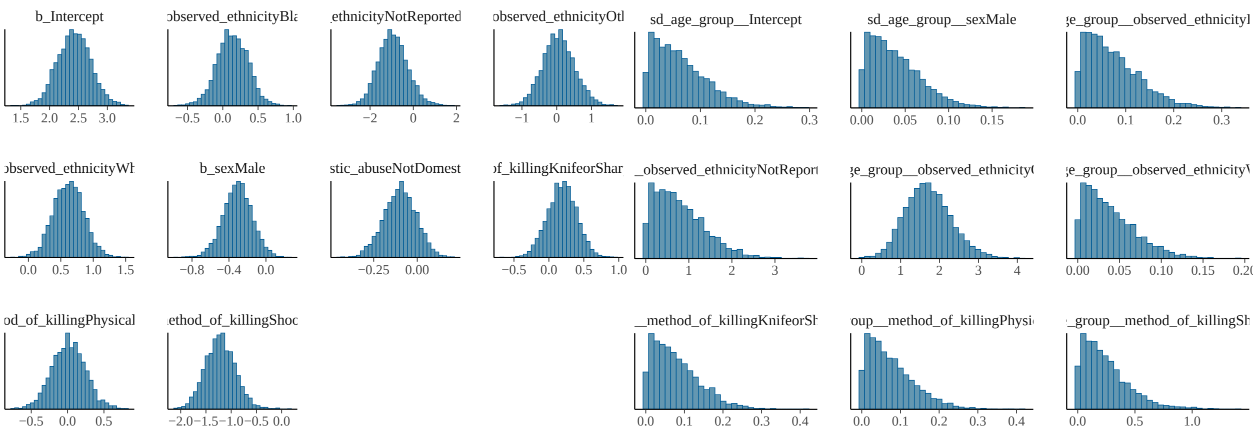
```
mcmc_plot(murder_fit, type = "hist", variable = "^b_", regex = TRUE)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
mcmc_plot(murder_fit, type = "hist", variable = "^sd_", regex = TRUE)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

ranef(murder_fit)

## $age_group
## , , Intercept
##
##                 Estimate  Est.Error       Q2.5     Q97.5
## 0 to 12      0.0055597031 0.07817563 -0.1578716 0.1795694
## 13 to 19     0.0165634441 0.07472933 -0.1226209 0.1997992
## 20 to 24    -0.0119298750 0.07353915 -0.1841728 0.1380567
## 25 to 34     0.0004566332 0.07052472 -0.1491251 0.1565034
## 35 to 44     0.0090539082 0.08277370 -0.1624511 0.2002284
## 45 to 54    -0.0116468960 0.07658333 -0.1912162 0.1358919
## 55 to 64    -0.0048140298 0.07860894 -0.1867454 0.1580583
## 65 and over -0.0003158503 0.07492299 -0.1708144 0.1583017
##
## , , sexMale
##
##                 Estimate  Est.Error        Q2.5      Q97.5
## 0 to 12     -0.0003753271 0.04914585 -0.11009449 0.10056151
## 13 to 19     0.0041952412 0.04689711 -0.09337021 0.10681818
## 20 to 24    -0.0076267652 0.04747644 -0.12086114 0.08508796
## 25 to 34    -0.0077053816 0.04773004 -0.12769364 0.08519591
## 35 to 44    -0.0013768313 0.05373352 -0.11795269 0.10930422
## 45 to 54    -0.0018896897 0.04638987 -0.10865953 0.09623448
## 55 to 64    -0.0034837550 0.04944297 -0.11486297 0.09928161
## 65 and over -0.0057688488 0.04773596 -0.11660165 0.08984264
##
## , , observed_ethnicityBlack
##
##                Estimate  Est.Error       Q2.5     Q97.5
## 0 to 12      0.003886841 0.09025751 -0.1897563 0.2082869
## 13 to 19     0.003309671 0.08185478 -0.1740733 0.1877554
## 20 to 24    -0.019585178 0.08426959 -0.2335110 0.1357573
## 25 to 34    -0.006842427 0.08278221 -0.2083864 0.1561353
## 35 to 44     0.025699463 0.10716575 -0.1711967 0.2825111
## 45 to 54    -0.008208156 0.09058968 -0.2151205 0.1668864
## 55 to 64     0.005708971 0.08973723 -0.1698931 0.2103880
## 65 and over -0.003892078 0.09161223 -0.2154061 0.1886334
##
## , , observed_ethnicityNotReportedDNotknown
##
##                Estimate Est.Error      Q2.5     Q97.5
## 0 to 12      0.001706759 1.0123765 -2.162581 2.2665543
## 13 to 19    -0.144153360 0.7785315 -2.006756 1.3945241
## 20 to 24    -0.542619826 1.0340669 -3.353260 0.9116282
## 25 to 34    -0.449848700 0.7496227 -2.286359 0.7147298
## 35 to 44     0.047442516 1.1080168 -2.397463 2.5340808
## 45 to 54     0.149461411 0.8749330 -1.506850 2.2715906
## 55 to 64     0.005075637 0.9676240 -2.037340 2.0657783
## 65 and over  0.312493169 0.8846644 -1.181598 2.4936211
```

```
##
## , , observed_ethnicityOther
##
##                  Estimate Est.Error      Q2.5      Q97.5
## 0 to 12        0.01759268 1.7895078 -3.779348  3.6813973
## 13 to 19      -0.07634373 0.9995035 -1.901640  2.0181077
## 20 to 24       1.01819454 1.4467117 -1.361432  4.2847766
## 25 to 34       1.17560704 1.3743514 -1.085076  4.2672629
## 35 to 44      -3.22071230 1.4378737 -6.466531 -0.7469167
## 45 to 54       0.88636918 1.4634257 -1.671045  4.2375164
## 55 to 64       0.41306112 1.6010852 -2.507193  3.9722232
## 65 and over    0.82159121 1.4318386 -1.576158  4.1197625
##
## , , observed_ethnicityWhite
##
##                    Estimate  Est.Error        Q2.5      Q97.5
## 0 to 12        3.534324e-04 0.04863553 -0.10608249 0.10786743
## 13 to 19       4.693516e-03 0.04890028 -0.08883494 0.12281349
## 20 to 24       6.580830e-03 0.04946130 -0.09179584 0.12573536
## 25 to 34       3.190070e-03 0.04718585 -0.09341796 0.10915453
## 35 to 44      -1.270665e-04 0.05674381 -0.12491400 0.12662162
## 45 to 54      -5.102433e-03 0.04758062 -0.11664347 0.09401067
## 55 to 64       4.129472e-05 0.05019295 -0.10474929 0.10665636
## 65 and over   -3.166475e-03 0.04893581 -0.11130256 0.09836965
##
## , , method_of_killingKnifeorSharpImplement
##
##                   Estimate  Est.Error       Q2.5     Q97.5
## 0 to 12        0.001880470 0.09755285 -0.2142410 0.2129810
## 13 to 19       0.032432147 0.09045773 -0.1158129 0.2673683
## 20 to 24      -0.029000480 0.09011422 -0.2544135 0.1272933
## 25 to 34      -0.020214911 0.08920682 -0.2462446 0.1490001
## 35 to 44       0.027136621 0.10177669 -0.1612859 0.2756120
## 45 to 54       0.002755345 0.09289251 -0.1959351 0.2073598
## 55 to 64      -0.002987762 0.09198914 -0.1979196 0.1922651
## 65 and over    0.013755572 0.09497156 -0.1663911 0.2399309
##
## , , method_of_killingPhysicalAssault
##
##                   Estimate  Est.Error       Q2.5     Q97.5
## 0 to 12        0.001036289 0.09281714 -0.1919212 0.2134907
## 13 to 19       0.002237163 0.08955034 -0.1904071 0.2025728
## 20 to 24       0.012359733 0.09533692 -0.1722983 0.2355050
## 25 to 34       0.015782786 0.09053661 -0.1573559 0.2327802
## 35 to 44      -0.000492217 0.11007684 -0.2482100 0.2367147
## 45 to 54      -0.008085729 0.09418839 -0.2272303 0.1872753
## 55 to 64      -0.016278174 0.09964132 -0.2645122 0.1728195
## 65 and over   -0.004966084 0.09781589 -0.2292755 0.1897025
##
## , , method_of_killingShooting
##
##                  Estimate Est.Error       Q2.5     Q97.5
## 0 to 12        0.02034245 0.3072568 -0.6430881 0.7198592
## 13 to 19      -0.10768870 0.2542669 -0.7408575 0.2909912
```

```
## 20 to 24      -0.07360012 0.2290926 -0.6562381 0.3117411
## 25 to 34       0.03570249 0.2133333 -0.4062624 0.4967139
## 35 to 44      -0.14758429 0.2860548 -0.8585297 0.2874381
## 45 to 54       0.10049562 0.2855391 -0.3971578 0.7923429
## 55 to 64       0.08100386 0.3127316 -0.4534174 0.8442592
## 65 and over    0.01638461 0.3086918 -0.6206460 0.7167116
```



First, looking at random effects, it appears the age group doesnt have much impact on the mean. Our estimates are all quite small and appear to be as likely to be positive and negative. The same can be said our sex Male, where the estimate do seem to be negative for ages above 18 but looking the at 95% confidence intervals, the model doesnt seem to be sure whether the effect is positive or negative. The effect of ethnicity Not Reported/Unknown seems to be higher for ages 20 to 35, our model seems confident the impact is negative i.e murder cases are less likely to be solved compared to the population average. This also appears to be the case for observed ethnicity Other, specifically for the 35 to 44 age group. Finally, method of killing shooting also seems to have a higher impact for ages 13 to 24. Let's look at the histogram of posterior distributions. We see that the intercept is centered around 2.4 which corresponds to about 91% and varies from 1.5 to 3.5 (88-97%). When looking at the observed ethnicity, as expected, we see that black is almost centered on 0 and has near equal chance to increase positively or negatively the mean. The same can be said for ethnicity Other. Ethnicity Not reported/Unknown however seems to high probability of negatively impacting the mean, being centered around -1. White also seems to have high probability of positively impacting the mean being centered around 0.5. Sex male also looks negative centered around -0.3, method of killing Knife seems positive and Physical Assault isn't clear. Shooting however seems very likely to be negative centered around -1.2. Looking at the standard deviation of our grouping effects, the plots seem consistent with the priors, they all seem to have moved away from mean 0, indicating an effect on the prior, most notably observed ethnicity Other. Next, we can compute these probabilities as Monte Carlo estimates.

```
samples = posterior_samples(murder_fit)

sum(samples$b_sexMale<0)/length(samples$b_sexMale)

## [1] 0.97625

sum(samples$b_observed_ethnicityBlack>0)/length(samples$b_observed_
    ethnicityBlack)

## [1] 0.73525

sum(samples$b_observed_ethnicityOther>0)/length(samples$b_observed_
    ethnicityOther)

## [1] 0.541
```

```r
sum(samples$b_observed_ethnicityWhite>0)/length(samples$b_observed_
    ethnicityWhite)
```

```
## [1] 0.9955
```

```r
sum(samples$b_method_of_killingShooting<0)/length(samples$b_method_of_
    killingShooting)
```

```
## [1] 0.99975
```

```r
sum(samples$b_method_of_killingPhysicalAssault>0)/length(samples$b_method_of_
    killingPhysicalAssault)
```

```
## [1] 0.52825
```

```r
sum(samples$b_method_of_killingKnifeorSharpImplement>0)/length(samples$b_
    method_of_killingKnifeorSharpImplement)
```

```
## [1] 0.818
```

```r
sum(samples$b_domestic_abuseNotDomesticAbuse<0)/length(samples$b_domestic_
    abuseNotDomesticAbuse)
```

```
## [1] 0.85225
```

As expected, we see that sex Male has 98% probability of negatively impacting the mean i.e reducing the probability of murder being solved compared to the population average, as well as method of killing shooting, which has probability close to 100%. We see that ethnicity White has 99% probability of positively impacting the mean. Interestingly, method of killing Knife also has 84% probability of positively impacting the mean, additionally, Not Domestic Abuse does seem to have a 85% probability of negatively impacting the mean. We see however that ethnicity Other and method of killing physical assault both have probabilities close to 50% so we are unsure how they affect the mean. Ethnicity Black has 73% chance of positively impacting the mean.

Monte Carlo Estimate and Error

In this section, we are generating 2 new "hypothetical homicides" during the Year after March. If we suppose the first homicide is A and the second is B, and let A,B be the events that the homicide A and B are solved. Here we will estimate :

$$P(A \wedge \tilde{B}|\text{data})$$

In this estimate, we will use all the minium number of samples needed from the posterior, w to get an MC error

$$<= 0.01$$

. If we wanted the minimum number of samples required, we can calculate that as so :

$$N > \hat{p}/(1-\hat{p})/0.01^2$$

```r
curated_cols <- c("recorded_date","age_group","sex","observed_ethnicity","
    domestic_abuse",
                  "borough","method_of_killing")
new_dates <- as.Date(c("2022-04-01", "2022-05-01", "2022-06-01",
                       "2022-07-01", "2022-08-01", "2022-09-01",
                       "2022-10-01", "2022-11-01", "2022-12-01"))
curated_homs <- dplyr::select(homicides, all_of(curated_cols))
```

```r
hypothetical_homicides <- tibble(recorded_date = sample(new_dates, 2, TRUE))
month_tibble <- read.csv("month_tibble.csv")
for(i in 2:length(names(curated_homs))){
  hypothetical_homicides <- cbind(hypothetical_homicides,
                                  sample(as.vector(unlist(unique(curated_homs
                                    [,i]))),2,TRUE))
}
names(hypothetical_homicides) <- names(curated_homs)
month_tibble$recorded_date = as.Date(month_tibble$recorded_date)
hypothetical_homicides <- as_tibble(hypothetical_homicides) %>%
  left_join(month_tibble, by = "recorded_date")
print(hypothetical_homicides)

## # A tibble: 2 x 10
##   recorded_~1 age_g~2 sex   obser~3 domes~4 borough metho~5  year month~6
    season
##   <date>      <chr>   <chr> <chr>   <chr>   <chr>   <chr>   <int> <chr>   <
    chr>
## 1 2022-05-01  55 to ~ Fema~ Asian   Not Do~ Hounsl~ Blunt ~    20 MAY
    spring
## 2 2022-10-01  20 to ~ Fema~ White   Domest~ Hackney Physic~    20 OCT
    autumn
## # ... with abbreviated variable names 1: recorded_date, 2: age_group,
## #   3: observed_ethnicity, 4: domestic_abuse, 5: method_of_killing,
## #   6: month.name

pred = predict(murder_fit, hypothetical_homicides, summary = FALSE)

MonteCarlo = function(N){
  prob = 0
  for (i in 1:N){
    if (pred[i,1] == 1 && pred[i,2] == 0){
      prob = prob + 1
    }
  }
  estimate = prob/N
  mcerror = sqrt(estimate*(1-estimate)/N)
  return(list(MCestimate = estimate, Error = mcerror))
}
N = 100
tMC = MonteCarlo(N)
while (tMC$Error > 0.01){
  N = N + 1
  tMC = MonteCarlo(N)
}
N

## [1] 367

tMC

## $MCestimate
## [1] 0.03814714
##
```

```
## $Error
## [1] 0.009998901
```

And so, our Monte Carlo estimate and error are as above, as well as the number of samples used.