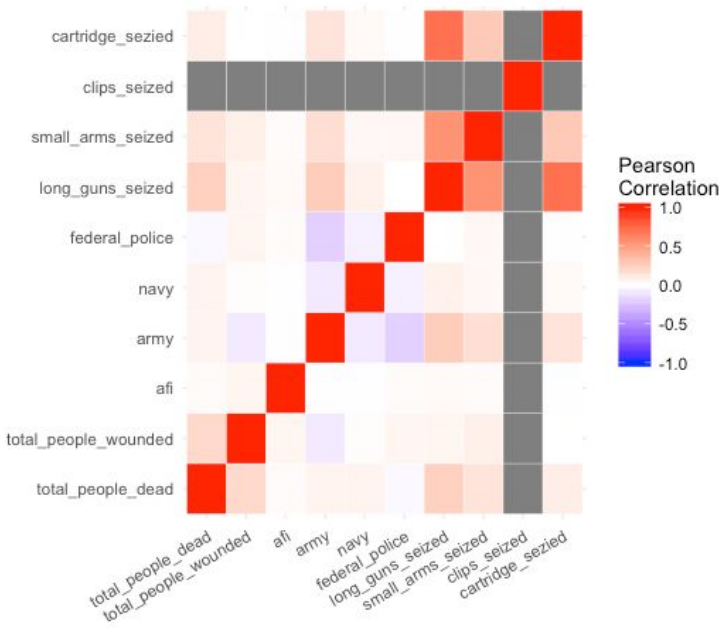
Data Challenge #2

Question #1

What factors lead to an increase in the total amount of people dead?

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.7693646  0.0412780  18.639  < 2e-16 ***
total_people_wounded 0.2358239  0.0163328  14.439  < 2e-16 ***
afi                  0.2792652  0.4955874   0.564  0.57312
army                 0.0297382  0.0628211   0.473  0.63596
navy                 0.4782041  0.1633872   2.927  0.00344 **
federal_police      -0.2871021  0.0908812  -3.159  0.00159 **
long_guns_seized     0.1518015  0.0101179  15.003  < 2e-16 ***
small_arms_seized   -0.0292220  0.0217092  -1.346  0.17834
clips_seized         0.0003087  0.0003707   0.833  0.40497
cartridge_sezied    -0.0001790  0.0000227  -7.884 3.82e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.035 on 5386 degrees of freedom
Multiple R-squared:  0.1065,  Adjusted R-squared:  0.105
F-statistic: 71.34 on 9 and 5386 DF,  p-value: < 2.2e-16
```



Conclusion: Involvement by the navy and the number of weapons seized, especially long guns, correlates with a higher amount of total deaths.

According to these summary statistics, the strongest correlation is with the navy and total dead (.48), but even so it is not very strong. Federal_police (-.29) and AFI (.28) involvement is slightly correlated as well.

The information featured in the correlation matrix reveals a different picture. It appears that the amount of weapons seized is related to the total number of people dead. We can assume that if more weapons, especially long arms, were taken in the event then more people likely died. However, this has to be truly explanatory because we cannot know the number of weapons seized ahead of the event. However, we can know which government group is deployed beforehand, so we can better estimate the deaths. This leads me to wonder how deployment is decided. Is there information available beforehand that leads the army, navy, federal police, etc to become involved in the event.

Question #2

Involvement by which law enforcement professionals will lead to more deaths?

Conclusion: We can predict if the navy or federal police are involved, the total number of deaths will rise.

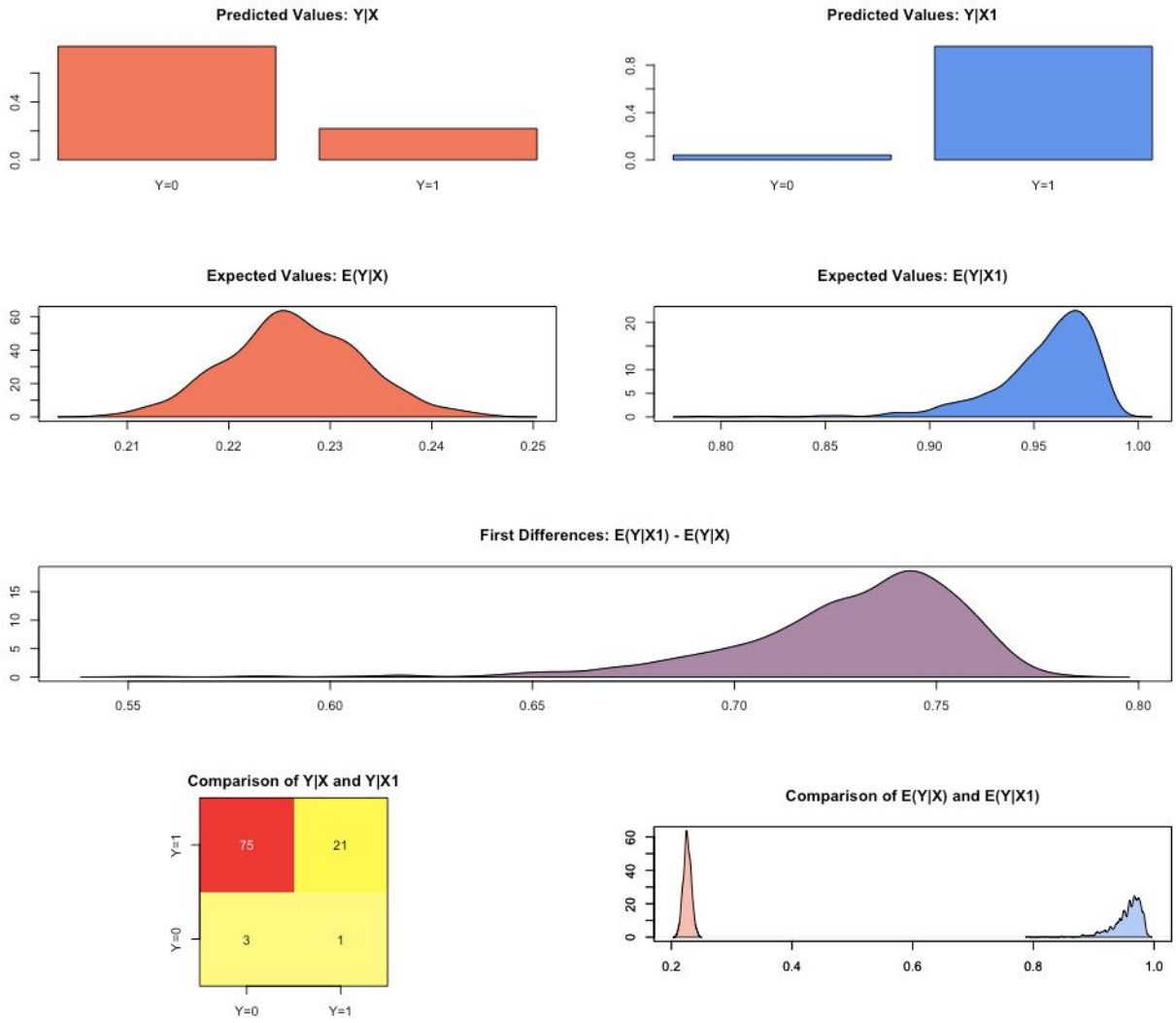
According to an assessment to identify the top features to include in the model, the navy and federal police involvement was positively correlated with total number of people killed in an event. We can assume from the numbers below that the involvement of either types of law enforcement will lead to a higher total of people dead from the event. The model also included cartridges and clips seized, but this cannot provide an applicable prediction because the amount of weapons seized is unknown before an event.

The AIC shows the best model includes 5 features: total_people_dead ~ total people_wounded + long_guns_seized + cartridge_sezied + federal_police + navy. The BIC instead included 4 features: total_people_dead ~ total people_wounded + long_guns_seized + cartridge_sezied + federal_police, excluding the navy.

Looking at the AIC model, we can confirm that both navy and federal police involvement leads to higher deaths in the event. We can also tell, using that model, that if everything else remains the same and the number of wounded increases to 30, the number of deaths will increase by 75%.

```
> #Top 10 models with low AIC
> subset_full$BestModels
   total_people_wounded long_guns_seized small_arms_seized cartridge_sezied clips_seized   afi  army federal_police navy
1                  TRUE             TRUE             FALSE             TRUE        FALSE FALSE FALSE           TRUE TRUE
2                  TRUE             TRUE              TRUE             TRUE        FALSE FALSE FALSE           TRUE TRUE
3                  TRUE             TRUE             FALSE             TRUE         TRUE FALSE FALSE           TRUE TRUE
4                  TRUE             TRUE              TRUE             TRUE         TRUE FALSE FALSE           TRUE TRUE
5                  TRUE             TRUE             FALSE             TRUE        FALSE  TRUE FALSE           TRUE TRUE
6                  TRUE             TRUE             FALSE             TRUE        FALSE FALSE  TRUE           TRUE TRUE
7                  TRUE             TRUE              TRUE             TRUE        FALSE  TRUE FALSE           TRUE TRUE
8                  TRUE             TRUE              TRUE             TRUE        FALSE FALSE  TRUE           TRUE TRUE
9                  TRUE             TRUE             FALSE             TRUE         TRUE  TRUE FALSE           TRUE TRUE
10                 TRUE             TRUE             FALSE             TRUE         TRUE FALSE  TRUE           TRUE TRUE
   Criterion
1   7671.142
2   7671.385
3   7672.458
4   7672.673
5   7672.827
6   7672.939
7   7673.082
8   7673.136
9   7674.131
10  7674.280
> #print best model
> subset_full$BestModel

Call:
lm(formula = y ~ ., data = data.frame(Xy[, c(bestset[-1], FALSE),
    drop = FALSE], y = y))

Coefficients:
        (Intercept)  total_people_wounded      long_guns_seized      cartridge_sezied        federal_police
           0.777992              0.233846              0.147948             -0.000173             -0.301210
                navy
           0.466117
```

```
> #Top 10 models with low BIC
> subset_full_bic$BestModels
   total_people_wounded long_guns_seized small_arms_seized cartridge_sezied clips_seized   afi  army federal_police
1                  TRUE             TRUE             FALSE             TRUE         FALSE FALSE FALSE           TRUE
2                  TRUE             TRUE             FALSE             TRUE         FALSE FALSE FALSE           TRUE
3                  TRUE             TRUE             FALSE             TRUE         FALSE FALSE FALSE          FALSE
4                  TRUE             TRUE             FALSE             TRUE         FALSE FALSE FALSE          FALSE
5                  TRUE             TRUE              TRUE             TRUE         FALSE FALSE FALSE           TRUE
6                  TRUE             TRUE              TRUE             TRUE         FALSE FALSE FALSE           TRUE
7                  TRUE             TRUE             FALSE             TRUE          TRUE FALSE FALSE           TRUE
8                  TRUE             TRUE             FALSE             TRUE          TRUE FALSE FALSE           TRUE
9                  TRUE             TRUE             FALSE             TRUE         FALSE  TRUE FALSE           TRUE
10                 TRUE             TRUE             FALSE             TRUE         FALSE FALSE  TRUE           TRUE
    navy Criterion
1  FALSE  7703.792
2   TRUE  7704.109
3   TRUE  7707.054
4  FALSE  7707.967
5  FALSE  7710.670
6   TRUE  7710.945
7  FALSE  7711.707
8   TRUE  7712.018
9  FALSE  7712.106
10 FALSE  7712.378
> #print best model
> subset_full_bic$BestModel

Call:
lm(formula = y ~ ., data = data.frame(Xy[, c(bestset[-1], FALSE),
    drop = FALSE], y = y))

Coefficients:
         (Intercept)  total_people_wounded      long_guns_seized      cartridge_sezied        federal_police
           0.7914243             0.2342176             0.1494172            -0.0001744            -0.3165588
```



Conclusion

We can see from the explanatory analysis that the total people dead is slightly positively correlated with the total people dead. This leads to the inference that in events with deaths the criminals have more weapons to use against law enforcement. Long arms has the strongest correlation versus the other weapons seized.

The involvement of the navy and AFI is linked to an increase in the total dead. This leads me to the question, why are these troops deployed to an event versus other law enforcement professionals? Are they used for events with more dangerous criminals or where more violence is projected?

We can predict that if there are more navy or federal police involved, the total dead will be higher. We need more information to know how to use this information. Should they equip these officials with more protection? Send more troops as reinforcements? Without more context, it is unclear how this prediction can be useful.

The models I used both in this report and in other analysis did not produce any significant results. At the same time, several tests I ran revealed these aren't the most accurate models (see code for these tests), so it would be important to run more tests to better understand the data. In addition, I would like to explore the time and location data as a predictive factor, since type of law enforcement involvement and weapons seized did not provide any significant conclusions.