Final Project

Jacob McCabe, Ben Bagley

12/3/2021

Research question(s).

The goal of this project is to create a model that predicts the number of casualties via gun violence in the United States and possibly predict the severity of an incident. Research question: Can we create a model that accurately predicts the number of casualties via gun violence in the United States?

Background and significance of the research.

According to evidence-based research from 2013, the US Gun Violence Archive estimates that the United States has seen over 40,000 gun violence deaths from 2013-2021. Approximately 19,000 of those deaths have been some form of homicide, and well over 20,000 have been via suicide. An additional 37,000 people have been injured by gun violence in that time as well.

As we continue to tackle the issue of gun violence in the United States – and with increased partisan interest in it over the last decade of presidential election cycles – it becomes imperative to be able to visualize the data and forecast it for the future. In order to have an educated conversation surrounding gun violence and as we set policies for the future, it is greatly beneficial to be able to estimate the statistics moving forward. With a steady rise in mass shootings and overall casualties, this research can help conceptualize the data in order to assist in the conversation about gun usage, ownership and policies in the United States.

The methods used to analyze and obtain the data.

In order to analyze the data, we wanted to start by getting a stationary series. If the original series is not stationary, we can perform differencing or the small trend method . This is crucial to perform latere analysis on the data. We also will be looking at the residuals to see whether or not they are normally distributed and/or independent and identically distributed (IID). This involves doing a Wilk-Shapiro test for normality, double checking with visual analysis of a QQ-plot, and a Portmanteau test. From here we will be able to look at the ACF and PACF plots of the stationary series to suggest possible models, and cross-validate the results by separating our data into a training and testing set and calculating the root mean square error (RMSE) for our potential models. Between the AIC from fitting the models to the original stationary series and the RMSE, we can use the model that makes the most sense to forecast.

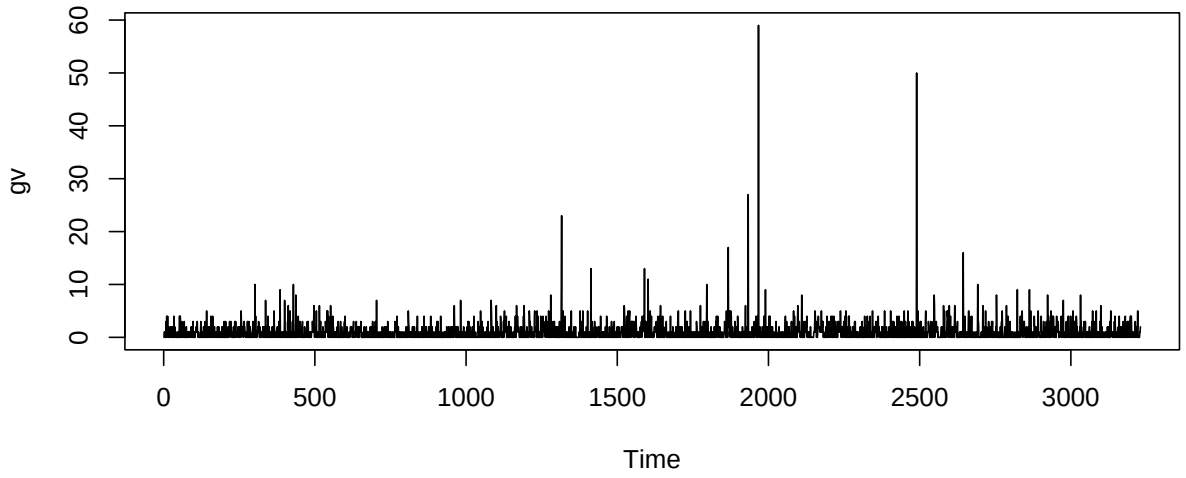
The results fo the analysis.

```
library(nortest)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
gv = read.csv("US-Gun-Violence.csv")
gv = ts(gv$killed)
```

```
plot(gv)
```

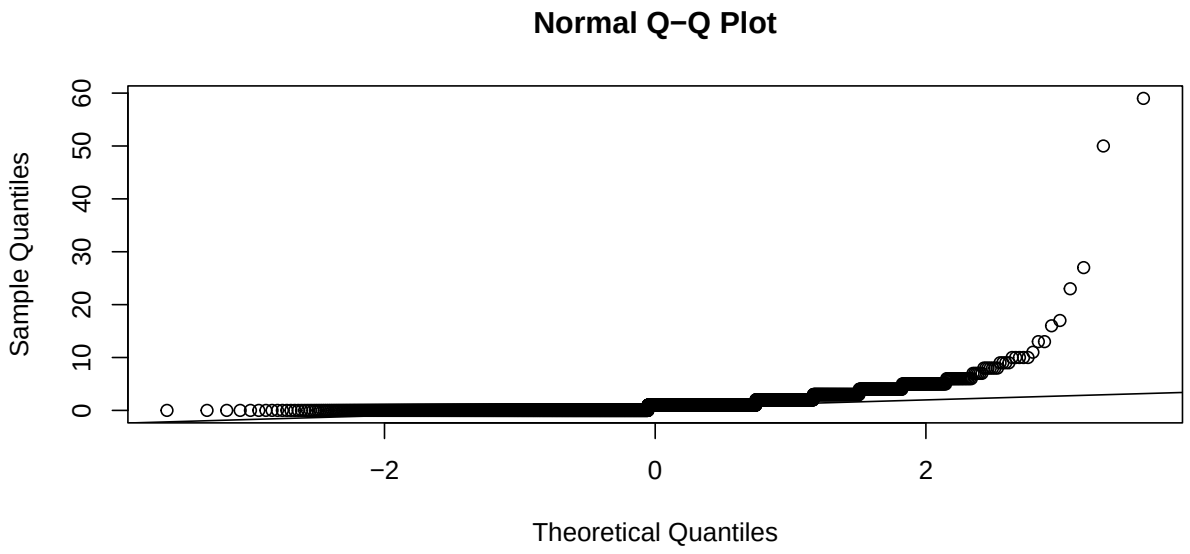


Looking at the plot of the number of people killed in each incident of gun violence, there appears to be a somewhat constant variance and potentially zero mean. From this we should be able to consider the time series to already be weakly stationary. Because it is already stationary, it would not be a good idea to use differencing or the small trend method since there is no seasonality or trend component to estimate in the model. Now we want to check the residuals of the model to see whether or not they can be considered to be approximately normally distributed and or IID.

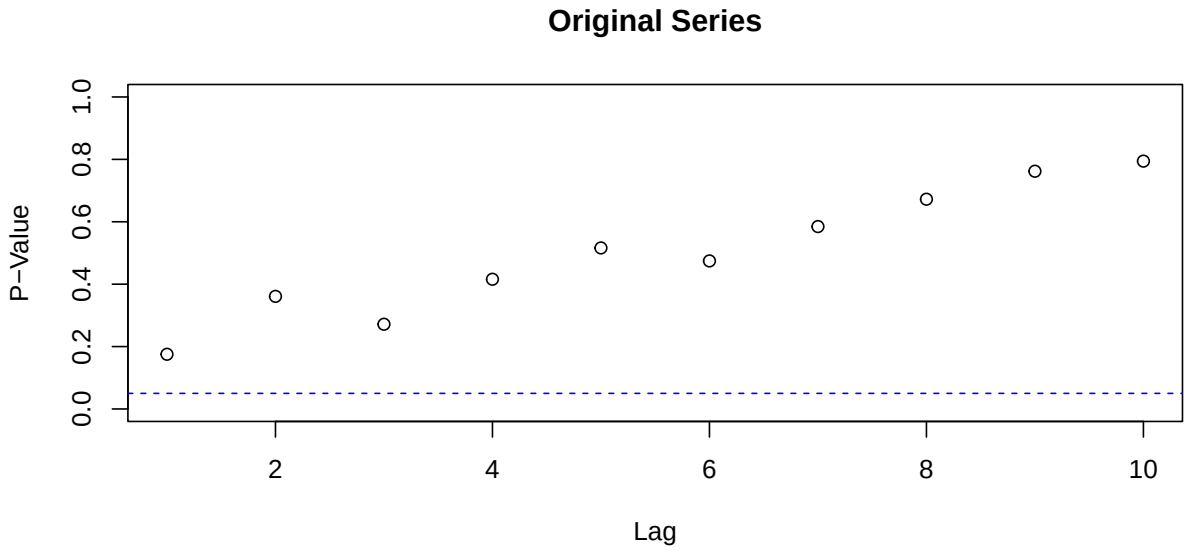
```
#Wilk-Shapiro test
ws_gv = shapiro.test(gv)
ws_gv
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gv
## W = 0.43158, p-value < 2.2e-16
```

```
#Q-Q plot
qqnorm(gv)
qqline(gv)
```



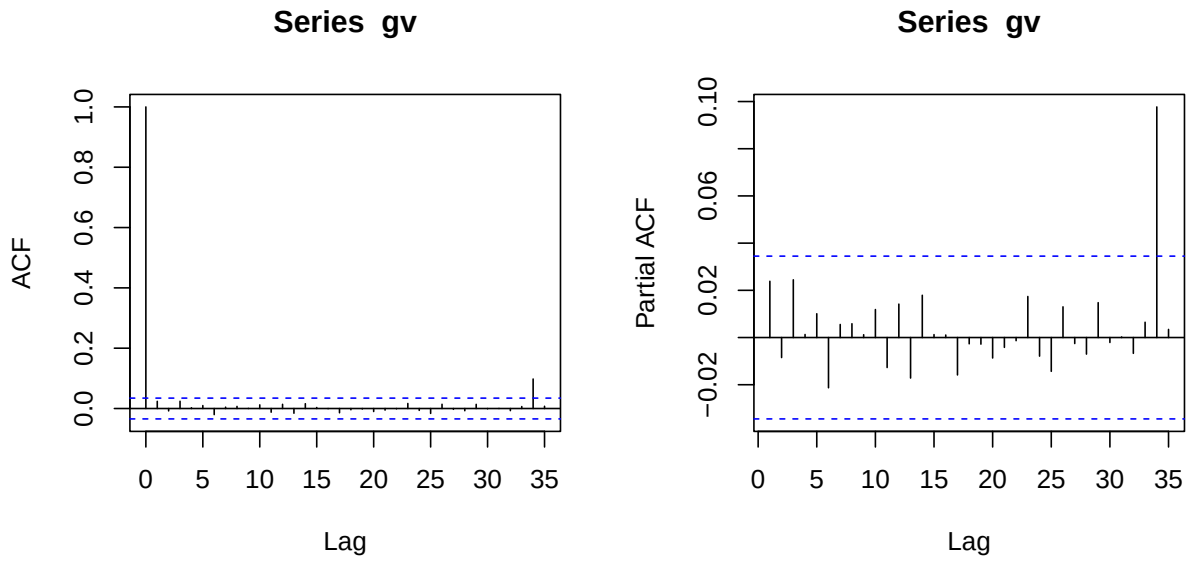

```
#Portmanteau test
pvalues_gv = double(10)
for(lag in 1:10){
  pvalues_gv[lag] = Box.test(gv, lag=lag)$p.value
}
plot(pvalues_gv, main="Original Series", xlab="Lag", ylab="P-Value",ylim=c(0,1))
abline(h=0.05, lty=2, col="blue")
```



Since the p-values from the Wilk-Shapiro test are all less than 2.2x10^{-16}, we reject the hypothesis that the residuals are approximately normally distributed. This is supported by visual analysis of the QQ plot. Since the p-values from the Portmanteau are all greater than 5%, we do not reject the hypothesis that the residuals are IID at a 95% confidence level.

```
par(mfrow = c(1,2))
acf(gv)
```

```
pacf(gv)
```



From looking at the ACF and PACF we see that the only significant lags are around lag 33, which we can say do not matter to the overall data since they are so far out. We can consider two possible models: ARMA(0,0) and ARMA(1,1). Only checking these models because the series appears to be white noise (ARMA(0,0)). We can try fitting the ARMA(1,1) simply because it can fit almost any series we try.

```
fit1 = arima(gv, order = c(0,0,0))
fit2 = arima(gv, order=c(1,0,1))
fit1
```

```
##
## Call:
## arima(x = gv, order = c(0, 0, 0))
##
## Coefficients:
##       intercept
##          1.0573
## s.e.     0.0366
##
## sigma^2 estimated as 4.329:  log likelihood = -6949.66,  aic = 13903.32
```

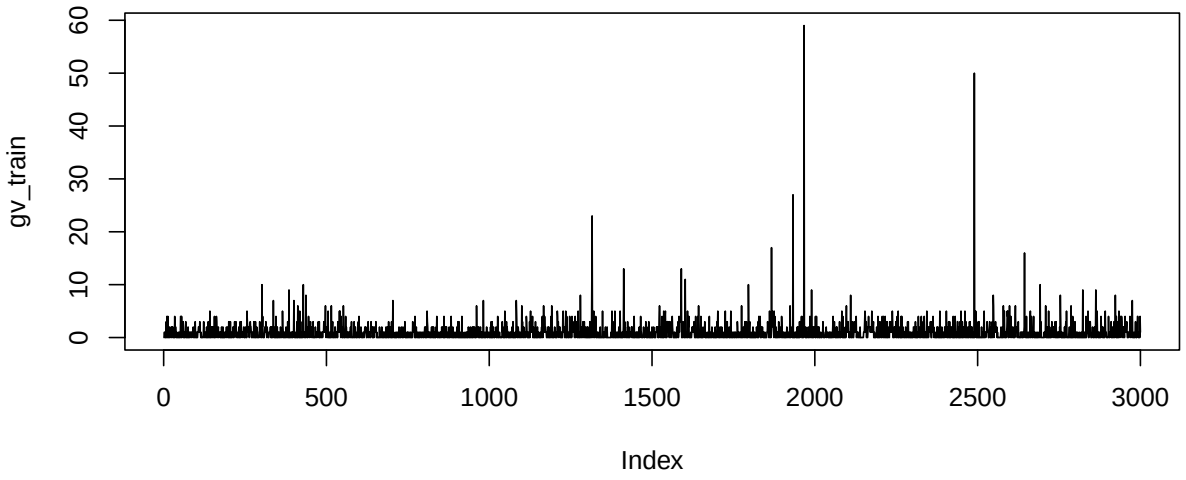
```
fit2
```

```
##
## Call:
## arima(x = gv, order = c(1, 0, 1))
##
## Coefficients:
##           ar1     ma1  intercept
##       -0.6396  0.6648     1.0577
## s.e.   0.2257  0.2199     0.0372
##
## sigma^2 estimated as 4.324:  log likelihood = -6947.92,  aic = 13903.84
```

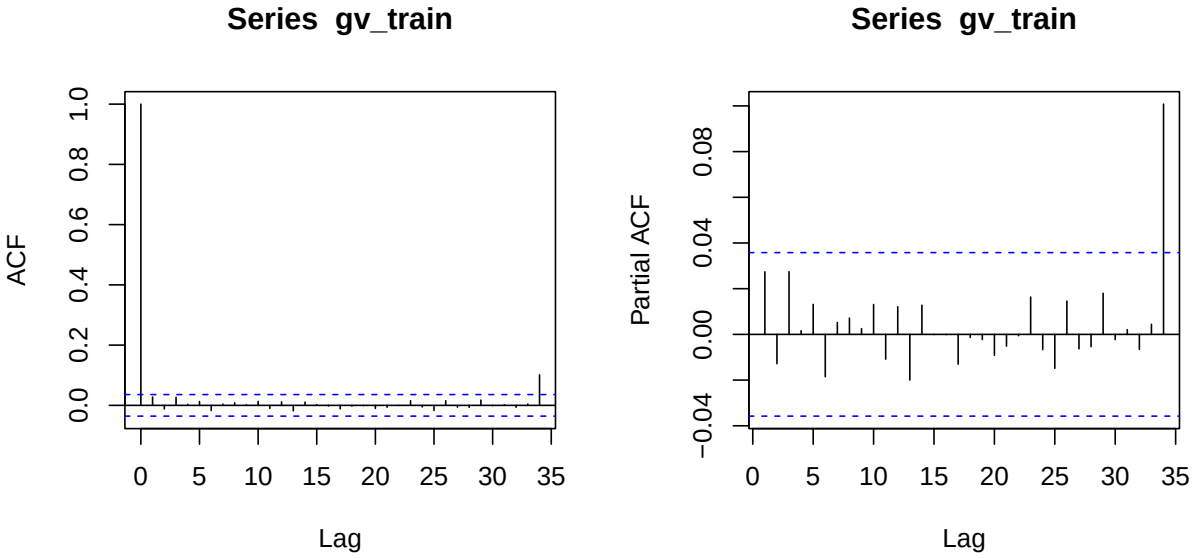
Since the AIC for the white noise model is slightly lower than the ARMA(1,1) model, in addition to having

fewer parameters to estimate, we will consider our model as white noise, i.e. ARMA(0,0). We can cross-validate this choice by separating our data into a training set and testing set and Checking the RMSE of each model.

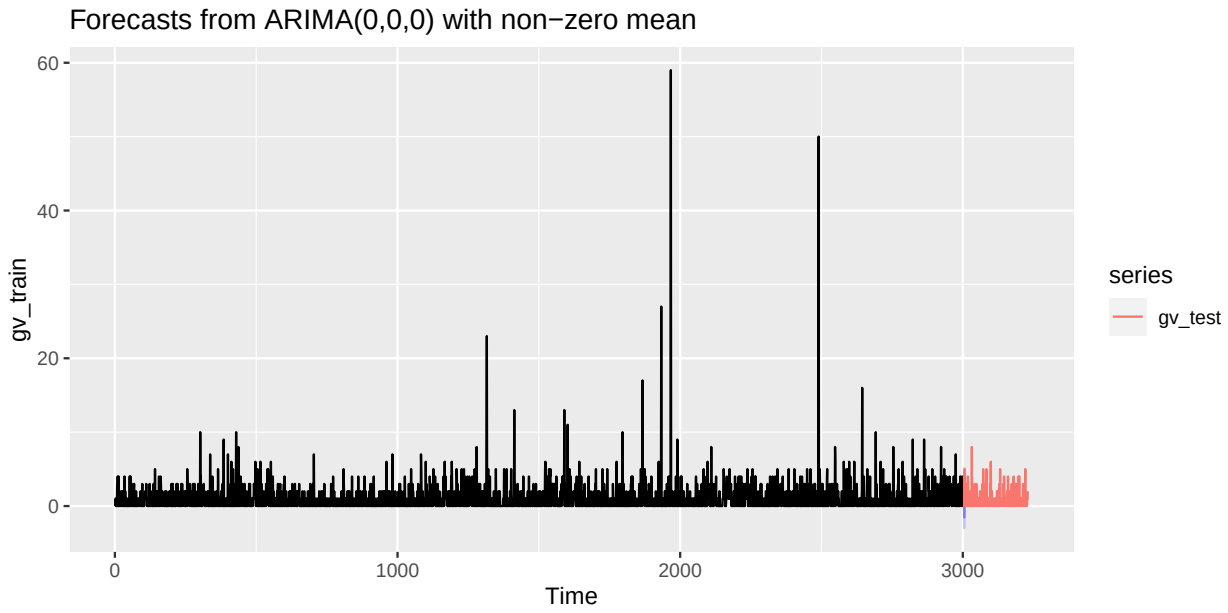
```
train = 1:3000
gv_train = gv[train]
test = 3001:3230
gv_test = ts(gv[test], start=3001)

plot (gv_train, type="l")
```




```
par(mfrow=c(1,2))
acf(gv_train)
pacf(gv_train)
```



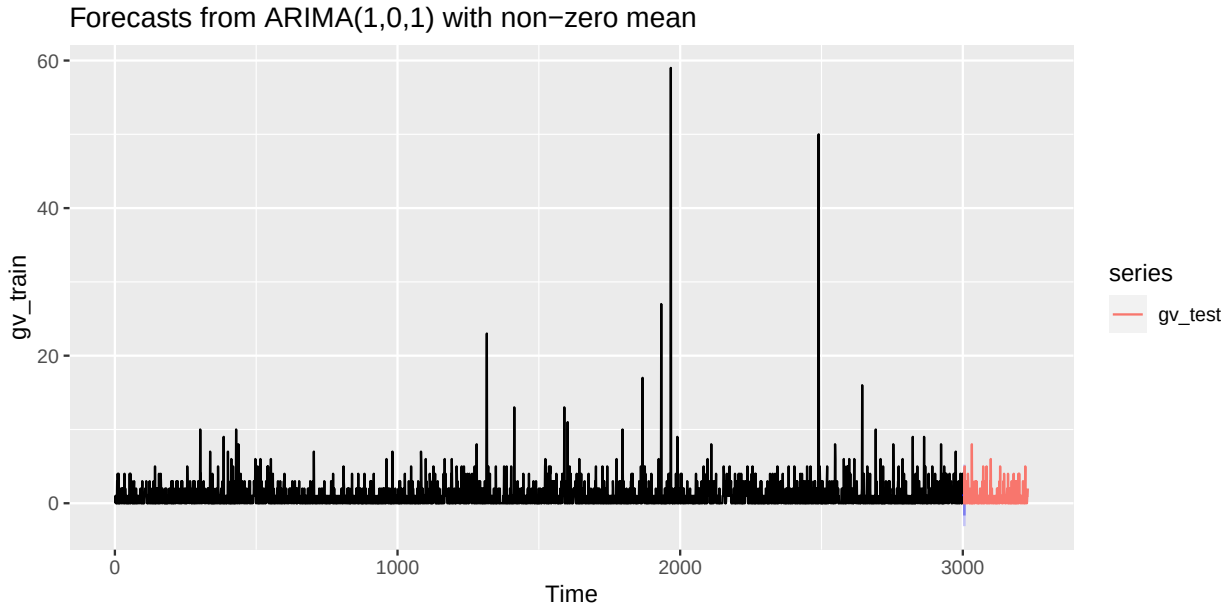
```
fit2 = arima(gv_train, order=c(0,0,0))
fit3 = arima(gv_train, order=c(1,0,1))

fit2_fore = forecast(fit2, 10)
autoplot(fit2_fore)+autolayer(gv_test)
```




```
sqrt(mean((gv_test - fit2_fore$mean)^2))
```

```
## [1] 1.790059
```

```
fit3_fore = forecast(fit3, 10)
autoplot(fit3_fore)+autolayer(gv_test)
```



```
sqrt(mean((gv_test - fit3_fore$mean)^2))
```

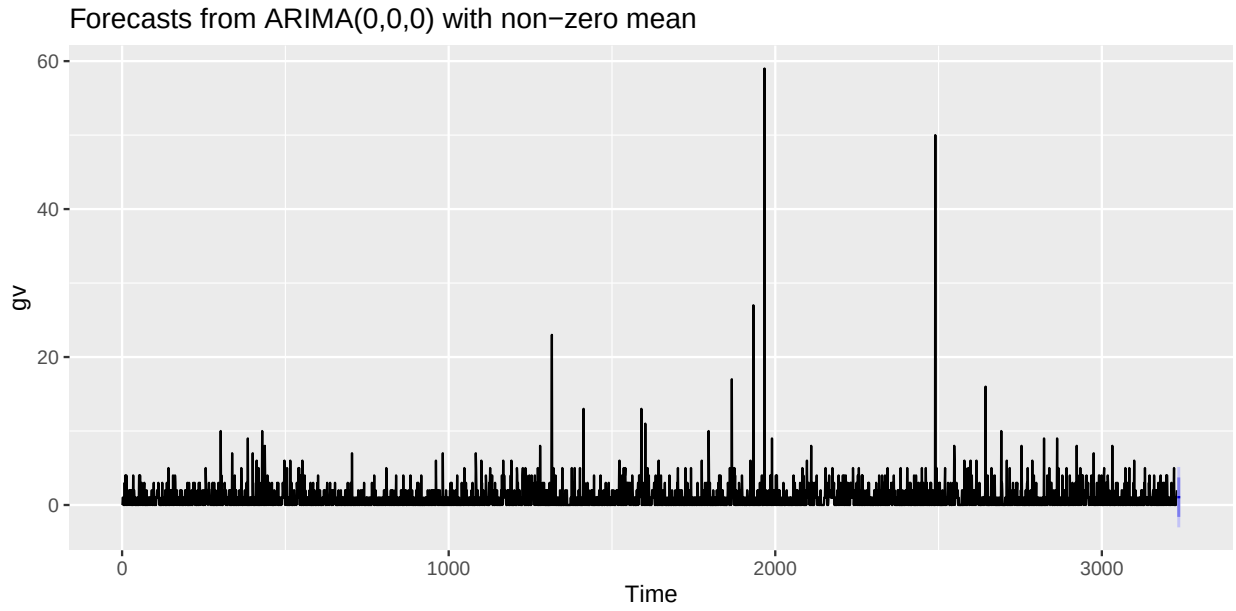
```
## [1] 1.787133
```

By using training data for both models, we get that the RMSE of the ARMA(0,0) and ARMA(1,1) models are approximately 1.790 and 1.787, respectively. Since the RMSE is virtually identical between the two models and we can continue with the ARMA(0,0) model due to it having fewer parameters to estimate.

```
fit1_fore = forecast(fit1, 10)
fit1_fore
```

```
##      Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## 3231       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3232       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3233       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3234       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3235       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3236       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3237       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3238       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3239       1.057276 -1.609127 3.723678 -3.020636 5.135187
## 3240       1.057276 -1.609127 3.723678 -3.020636 5.135187
```

```
autoplot(fit1_fore)
```



The results of the forecasting seem to show that the next 10 incidents of gun violence will likely only have one person being killed in each incident. Since the model is white noise, it is simply predicting the mean of the series for each forecasted incident. It is inherently difficult to forecast the severity of incidents with this model.

A discussion of the research, the limitations of the current research, any assumptions made, and possibilities of future work that should be conducted.

The research shows a haunting yet understandable conclusion: gun violence in the United States largely happens in random occurrences. There is no seasonality, which means there is no specific month or time of the year that gun violence most occurs. There is no trend, meaning that gun violence has remained steadily

present over the seven years in the data set. With the final selected model being just white noise, the data does not present that there is any way to accurately predict gun violence in the country.

This research did not contain different kinds of gun violence, which could affect the model if observed independently. There are different statistics for suicide, homicide, mass shootings, accidental deaths and more, and breaking up the data into these individual sets could provide a more holistic view at the problem. Also, as we continue to gather data, a data set spanning more years could yield stronger results.

When discussing the results in an application to reality, a white noise model provides little to the conversation. The data is in no way useless, but it doesn't provide lawmakers or concerned parties with any substantial conclusions, it doesn't further the conversation. As it stands, the best model to predict gun violence in the United States is a random one, indicating the severity of the issue in the country.

References

https://www.kaggle.com/konivat/us-gun-violence-archive-2014

https://www.gunviolencearchive.org/