R Camp 2019 HW1

Alex Pugh

8/19/2019

Problem 1

The Correlates of War project collects data on wars and militarized disputes. One of the datasets that they have focuses on inter-state wars. I am interested in factors that impact the number of battle deaths within interstate wars. My first research question addresses whether war initiators experience fewer battle deaths than their targets. I expect that states that initiate war have considered the costs and benefits of war and believe that they are likely to beat their opponent. As a result, they might be more prepared for war than their opponent and choose opponents that are militarily weaker. The lack of preparedness and disadvantage of the target are more likely to result in higher battle deaths compared to the initiator.

My second question concerns the relation between time and battle deaths. Has the number of battle deaths in a war increased or decreased over time? War fighting technology has improved dramatically, but these technologies have enabled countries to fight with fewer soldiers directly in combat. Also, technology has been used to better protect soldiers in combat. As a result, I expect that the number of battle deaths has generally decreased over time. The world wars might skew the data so I will explore with and without the world wars if they appear to be outliers.

Data Inspection

First, we need to load the datset and inspect the data, preforming data cleaning where needed.

```
knitr::opts_chunk$set(echo = TRUE)
# clear environment
rm(list=ls())

# set wd
setwd("/Users/akg6//Dropbox/R Camp/Homework 1") #school computer
#setwd("/Users/alexgoodman//Dropbox/R Camp/Homework 1") #home computer
getwd()
```

```
## [1] "/Users/akg6/Dropbox/R Camp/Homework 1"
```

```
# load COW interstate war data
dat <- read.csv("Inter-StateWarData_v4.0.csv", header=T, sep=",")

#see names
names(dat)
```

```
##  [1] "WarNum"      "WarName"     "WarType"     "ccode"       "StateName"
##  [6] "Side"        "StartMonth1" "StartDay1"   "StartYear1"  "EndMonth1"
## [11] "EndDay1"     "EndYear1"    "StartMonth2" "StartDay2"   "StartYear2"
## [16] "EndMonth2"   "EndDay2"     "EndYear2"    "TransFrom"   "WhereFought"
## [21] "Initiator"   "Outcome"     "TransTo"     "BatDeath"    "Version"
```

```r
#rename battle deaths variable
names(dat)[names(dat)=="BatDeath"] <- "Battle Deaths"

# verify name change
names(dat)
```

```
##  [1] "WarNum"        "WarName"       "WarType"       "ccode"
##  [5] "StateName"     "Side"          "StartMonth1"   "StartDay1"
##  [9] "StartYear1"    "EndMonth1"     "EndDay1"       "EndYear1"
## [13] "StartMonth2"   "StartDay2"     "StartYear2"    "EndMonth2"
## [17] "EndDay2"       "EndYear2"      "TransFrom"     "WhereFought"
## [21] "Initiator"     "Outcome"       "TransTo"       "Battle Deaths"
## [25] "Version"
```

```r
#determine class of initiator variable
class(dat$Initiator)
```

```
## [1] "integer"
```

```r
#create factor variable for intiator
dat$Initiate <- factor(dat$Initiator, labels=c("Initiator","Target"))

#determine class of new variable-should be factor
class(dat$Initiate)
```

```
## [1] "factor"
```

```r
#subset data, exclude month and day variables as well as war tranformed variables
dat2 <- dat[,c(1:6, 9, 12, 15, 18, 21:22, 24:26)]

#check first lines to ensure worked
head(dat2, n=2)
```

```
##   WarNum            WarName WarType ccode StateName Side StartYear1
## 1      1 Franco-Spanish War       1   230     Spain    2       1823
## 2      1 Franco-Spanish War       1   220    France    1       1823
##   EndYear1 StartYear2 EndYear2 Initiator Outcome Battle Deaths Version
## 1     1823         -8       -8         2       2           600       4
## 2     1823         -8       -8         1       1           400       4
##    Initiate
## 1    Target
## 2 Initiator
```

```r
#save as rdata file
save(dat2, file = "cow.RData")
```

Descriptive Statistics

Now that the dataset has been cleaned, we can look at some descriptive statistics.

```r
# clear environment
rm(list=ls())
#load data
load("cow.RData")
# load stargazer package
library("stargazer")
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
# create Summary Statistics
stargazer(dat2, type = "text", title="Summary Statistics", median=TRUE)
```

```
##
## Summary Statistics
## ================================================================================
## Statistic       N      Mean      St. Dev.     Min  Pctl(25) Median Pctl(75)    Max
## --------------------------------------------------------------------------------
## WarNum         337   126.920     61.237       1      82     139     172       227
## WarType        337    1.000       0.000       1      1       1       1         1
## ccode          337   419.908     246.288      2     220     355     652       920
## Side           337    1.418       0.494       1      1       1       2         2
## StartYear1     337 1,930.573     46.608    1,823  1,900   1,939   1,969     2,003
## EndYear1       337 1,931.828     46.704    1,823  1,900   1,941   1,973     2,003
## StartYear2     337   100.282     443.783     -8      -8      -8      -8      1,974
## EndYear2       337   100.291     443.820     -8      -8      -8      -8      1,974
## Initiator      337    1.677       0.468       1      1       2       2         2
## Outcome        337    2.092       1.528       1      1       2       2         8
## Battle Deaths 337 95,195.550 501,199.700  -9     400    2,000   10,000  7,500,000
## Version        337    4.000       0.000       4      4       4       4         4
## --------------------------------------------------------------------------------
```

```r
# Battles Deaths has a min of -9
# -9 means that the data is unknown according to the codebook
# recode -9 as NA
dat2$`Battle Deaths`[dat2$`Battle Deaths`==-9] <- NA

#save as rdata file
save(dat2, file = "cow.RData")

# recreate Summary Statistics
stargazer(dat2, type = "text", title="Summary Statistics", median=TRUE)
```

```
##
## Summary Statistics
## ================================================================================
## Statistic       N      Mean      St. Dev.     Min  Pctl(25)  Median    Pctl(75)       Max
```

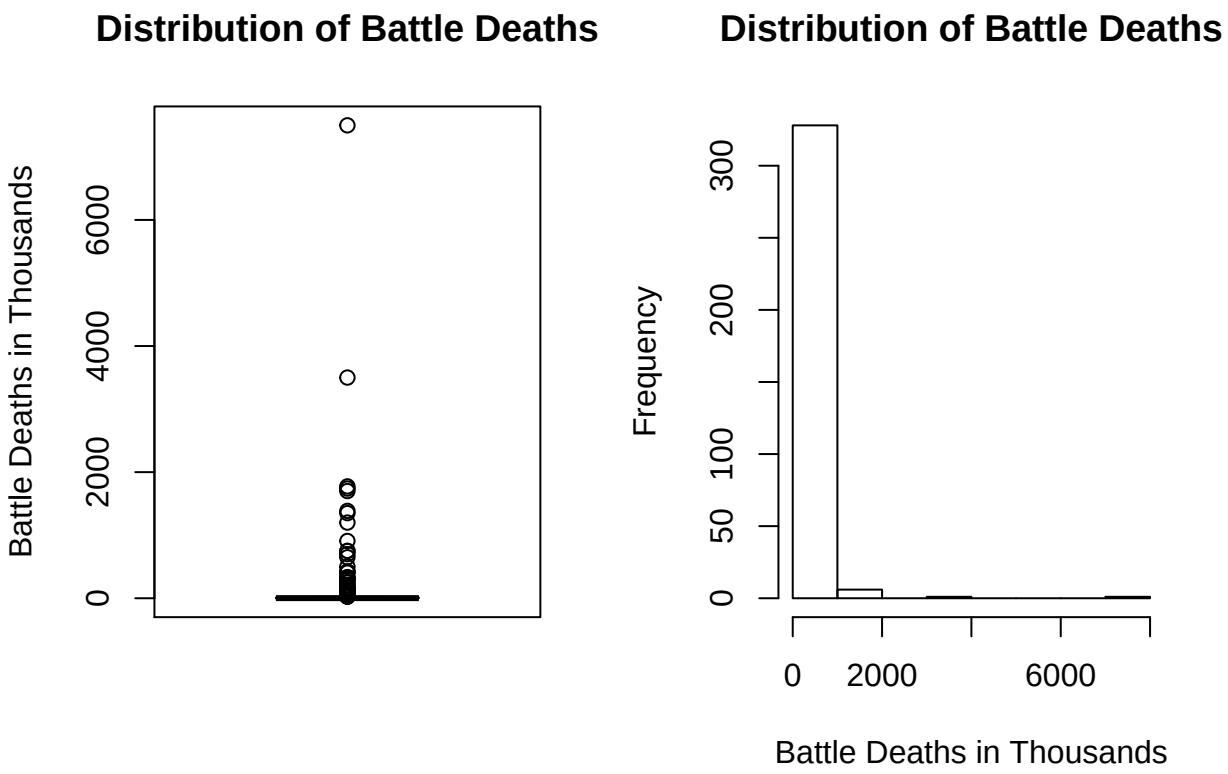
```
## ---------------------------------------------------------------------------------------
## WarNum         337   126.920      61.237       1       82        139        172          227
## WarType        337    1.000        0.000       1       1          1          1            1
## ccode          337   419.908      246.288      2      220        355        652          920
## Side           337    1.418        0.494       1       1          1          2            2
## StartYear1     337 1,930.573     46.608     1,823   1,900      1,939      1,969        2,003
## EndYear1       337 1,931.828     46.704     1,823   1,900      1,941      1,973        2,003
## StartYear2     337   100.282      443.783     -8      -8         -8         -8          1,974
## EndYear2       337   100.291      443.820     -8      -8         -8         -8          1,974
## Initiator      337    1.677        0.468       1       1          2          2            2
## Outcome        337    2.092        1.528       1       1          2          2            8
## Battle Deaths 336 95,478.900 501,920.200 0.000 400.000  2,000.000 10,000.000 7,500,000.000
## Version        337    4.000        0.000       4       4          4          4            4
## ---------------------------------------------------------------------------------------
```

From this table, we can see that the temporal range of our data extends from 1823-2003. As one would expect, the range of battle deaths is quite large. Interestingly, the minimum number of battle deaths is zero. In the dataset, for the state to be considered a war participant, it must have committed atleast 1,000 troops to the war or suffered atleast 100 battle related deaths. This means that there are a few states that committed atleast 1,000 troops and did not suffer any battle casualties. A brief inspection of the data indicates that most of the cases of 0 troop deaths occured in the late 1990s or early 2000s, but we will look into that more later. First, we will explore the distribution of battle deaths using a histograms and box plots. Because there is such a large range in our number of battle deaths, we will create a new variable that is battle deaths in thousands in order to make clearer plots.

```r
# create new variable of battle deaths in thousands
dat2$Deaths1000 <- dat2$`Battle Deaths`/1000
head(dat2, n=2)
```

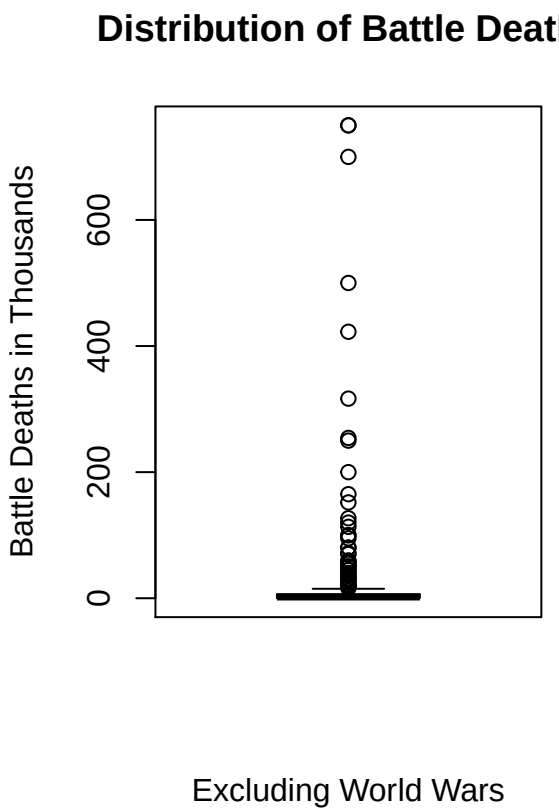
```
##   WarNum            WarName WarType ccode StateName Side StartYear1
## 1      1 Franco-Spanish War       1   230     Spain    2       1823
## 2      1 Franco-Spanish War       1   220    France    1       1823
##   EndYear1 StartYear2 EndYear2 Initiator Outcome Battle Deaths Version
## 1     1823         -8       -8         2       2           600       4
## 2     1823         -8       -8         1       1           400       4
##    Initiate Deaths1000
## 1    Target        0.6
## 2 Initiator        0.4
```

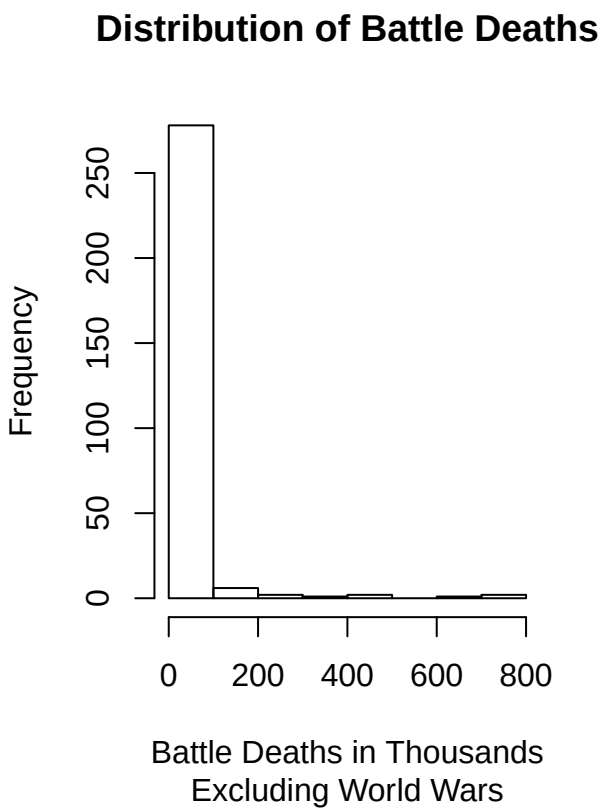
```r
#create boxplot and histogram of distribution
par(mfrow=c(1,2))
boxplot(dat2$Deaths1000, main="Distribution of Battle Deaths",
        ylab="Battle Deaths in Thousands")
hist(dat2$Deaths1000, main="Distribution of Battle Deaths",
     xlab="Battle Deaths in Thousands" )
```



From these graphs, we can see that we have a few outliers with a high number of battle deaths, with most states incurring less than 1,000 battle deaths. To get a better idea of the distribution, we will remove the two world wars from the dataset as they are the likely sources of the outliers.

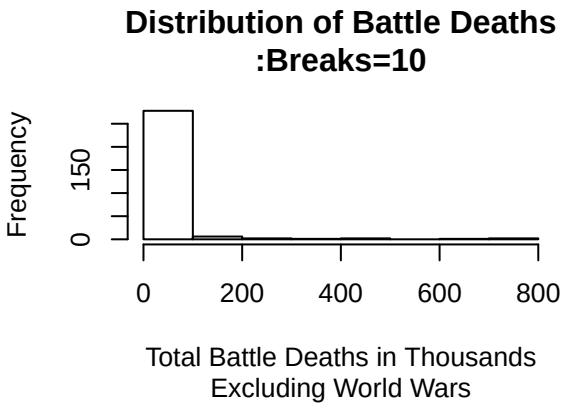
```
# create new dataframe without World War values
datnww <- dat2[dat2$WarNum != 106,]
datnww <- datnww[datnww$WarNum != 139,]

#View(datnww)
# create dataframe of only World War values
datww <- dat2[dat2$WarNum == 106|dat2$WarNum == 139,]

#View(datww)

#create boxplot and histogram of distribution
par(mfrow=c(1,2))
boxplot(datnww$Deaths1000, main="Distribution of Battle Deaths",
        sub="Excluding World Wars", ylab="Battle Deaths in Thousands")
hist(datnww$Deaths1000, main="Distribution of Battle Deaths",
     sub="Excluding World Wars", xlab="Battle Deaths in Thousands" )
```

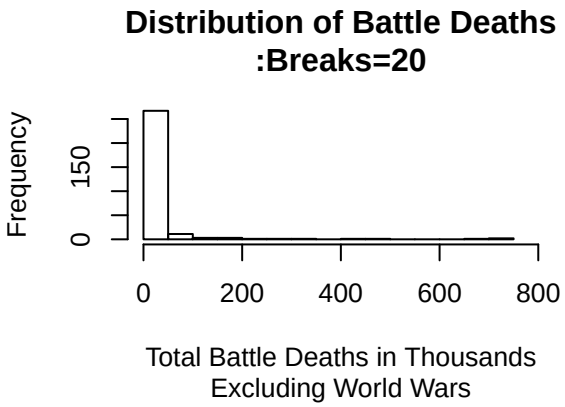


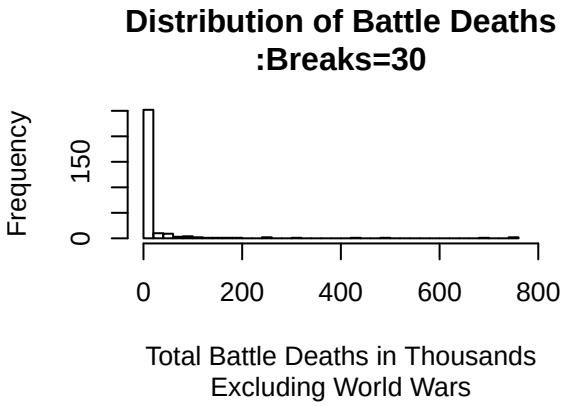


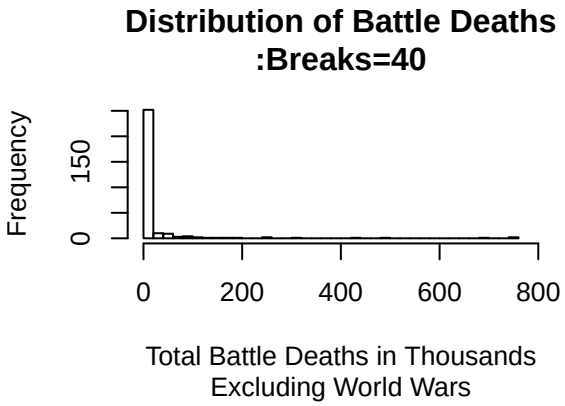
We can change the number of breaks in the histogram to get a better look at the distribution of battle deaths. There is still a great deal of variation in battle deaths, even with the exclusion of the World Wars.

```
#create histograms, adjusting number of bins
par(mfrow=c(2,2))
hist(datnww$Deaths1000, breaks= 10, main="Distribution of Battle Deaths\n:Breaks=10",
     sub="Excluding World Wars", xlab="Total Battle Deaths in Thousands" )
hist(datnww$Deaths1000, breaks= 20, main="Distribution of Battle Deaths\n:Breaks=20",
     sub="Excluding World Wars", xlab="Total Battle Deaths in Thousands", xlim = c(0, 800))
hist(datnww$Deaths1000, breaks= 30, main="Distribution of Battle Deaths\n:Breaks=30",
     sub="Excluding World Wars", xlab="Total Battle Deaths in Thousands" , xlim = c(0, 800))
hist(datnww$Deaths1000, breaks= 40, main="Distribution of Battle Deaths\n:Breaks=40",
     sub="Excluding World Wars", xlab="Total Battle Deaths in Thousands" , xlim = c(0, 800))
```










```
#summary statistics of not world wars battle deaths variable
summary(datnww$`Battle Deaths`)
```

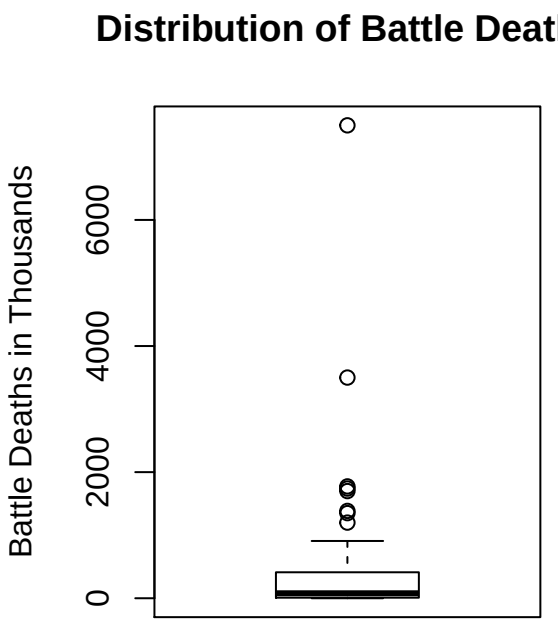
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     300    1200   23520    6325  750000       1
```

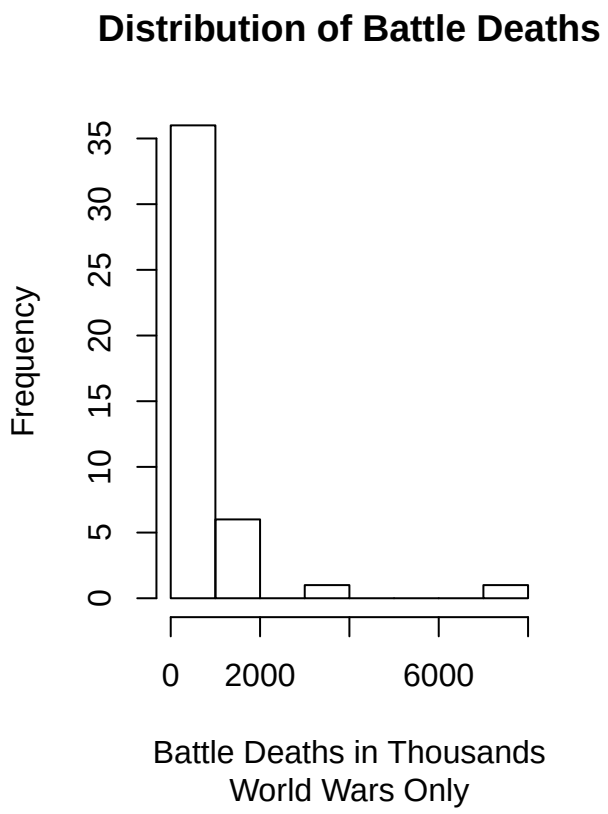
```
#standrad deviation of not world wars battle deaths variable
sd(datnww$`Battle Deaths`, na.rm = T)
```

```
## [1] 89593.41
```

We can also look at the variation in battle deaths amonst those that participated in the World Wars. There appears to be a great deal of variation, with a minimum of 300 and a maximum of 7,500,000.

```
#create boxplot and histogram of distribution
par(mfrow=c(1,2))
boxplot(datww$Deaths1000, main="Distribution of Battle Deaths",
        sub="World Wars Only", ylab="Battle Deaths in Thousands")
hist(datww$Deaths1000, main="Distribution of Battle Deaths",
     sub="World Wars Only", xlab="Battle Deaths in Thousands" )
```






```
#summary statistics of world wars battle deaths variable
summary(datww$`Battle Deaths`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     300    8925   78750  573021  408741 7500000
```

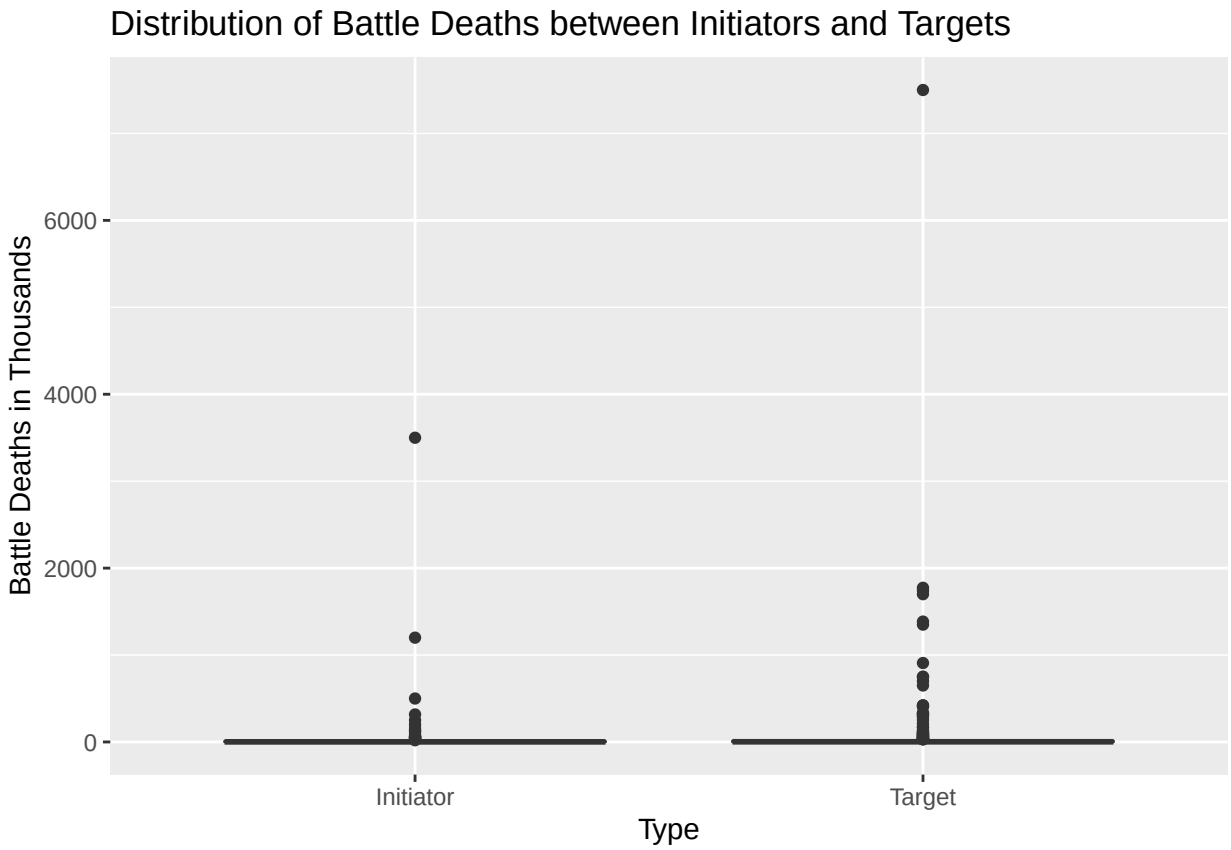
```
#standrad deviation of world wars battle deaths variable
sd(datww$`Battle Deaths`, na.rm = T)
```

```
## [1] 1280558
```

Now that we understand the distribution of battle deaths, we can begin to the relationship between battle deaths and other variables, specifically intitation and time. First, we can plot the distribution of battle deaths across those that initated the war and those that were targets. Both initators and targets have a few outliers that make it difficult to see if there is much difference in the distribution of battle deaths. To get a better look at the distribution, we can exclude outliers by limiting the data to battle deaths less than or equal to 12,000.

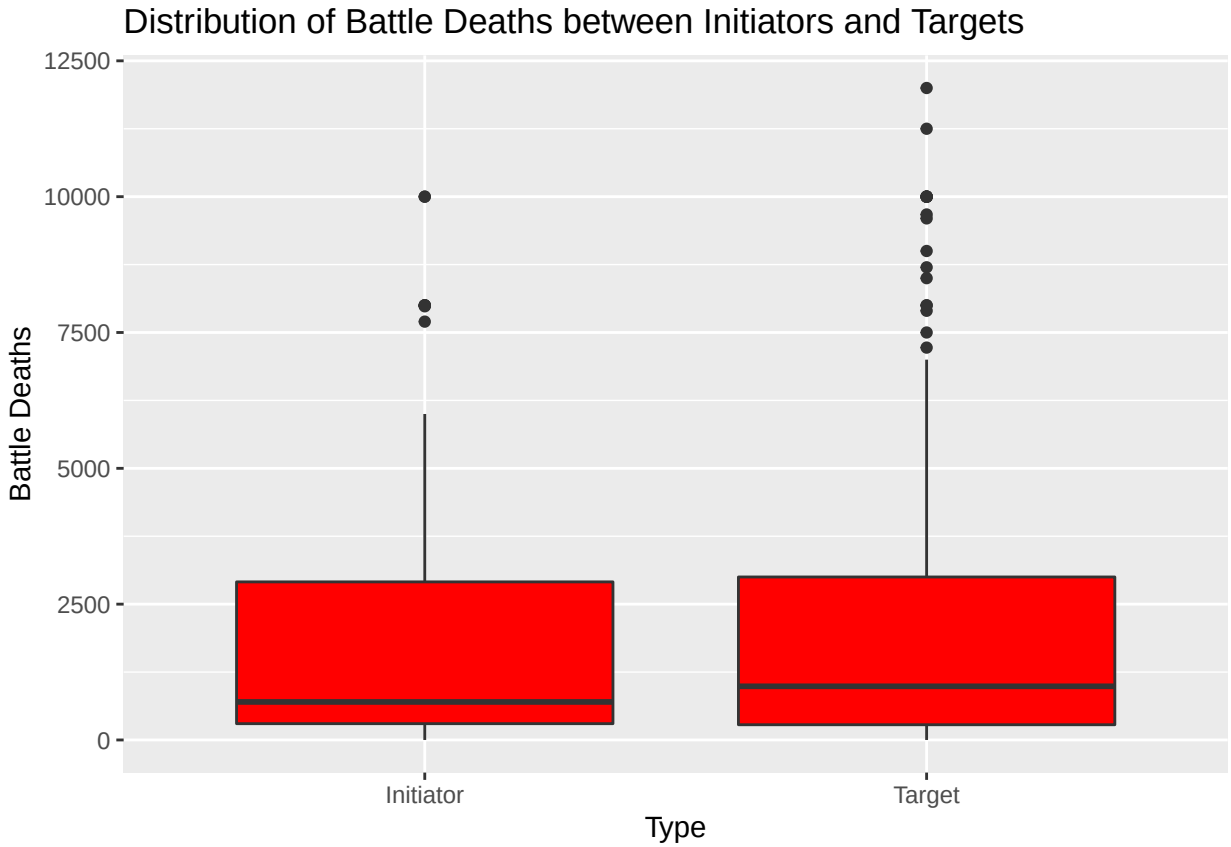
```
#load ggplot2 package
library(ggplot2)
#create boxplot of battledeaths in thousands by type using ggplot2
ggplot(dat2, aes(x=Initiate, y=Deaths1000)) + geom_boxplot(fill="red") + ggtitle("Distribution of Battl
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```




```
#create a new variable that is missing for any battle deaths greater than 12000
dat2$Death12000 <- ifelse(dat2$`Battle Deaths` <= 12000, dat2$`Battle Deaths`, NA)

#create boxplot of battledeaths less than 12000 by type using ggplot2
ggplot(dat2, aes(x=Initiate, y=Death12000)) + geom_boxplot(fill="red") + ggtitle("Distribution of Battl
```

```
## Warning: Removed 82 rows containing non-finite values (stat_boxplot).
```

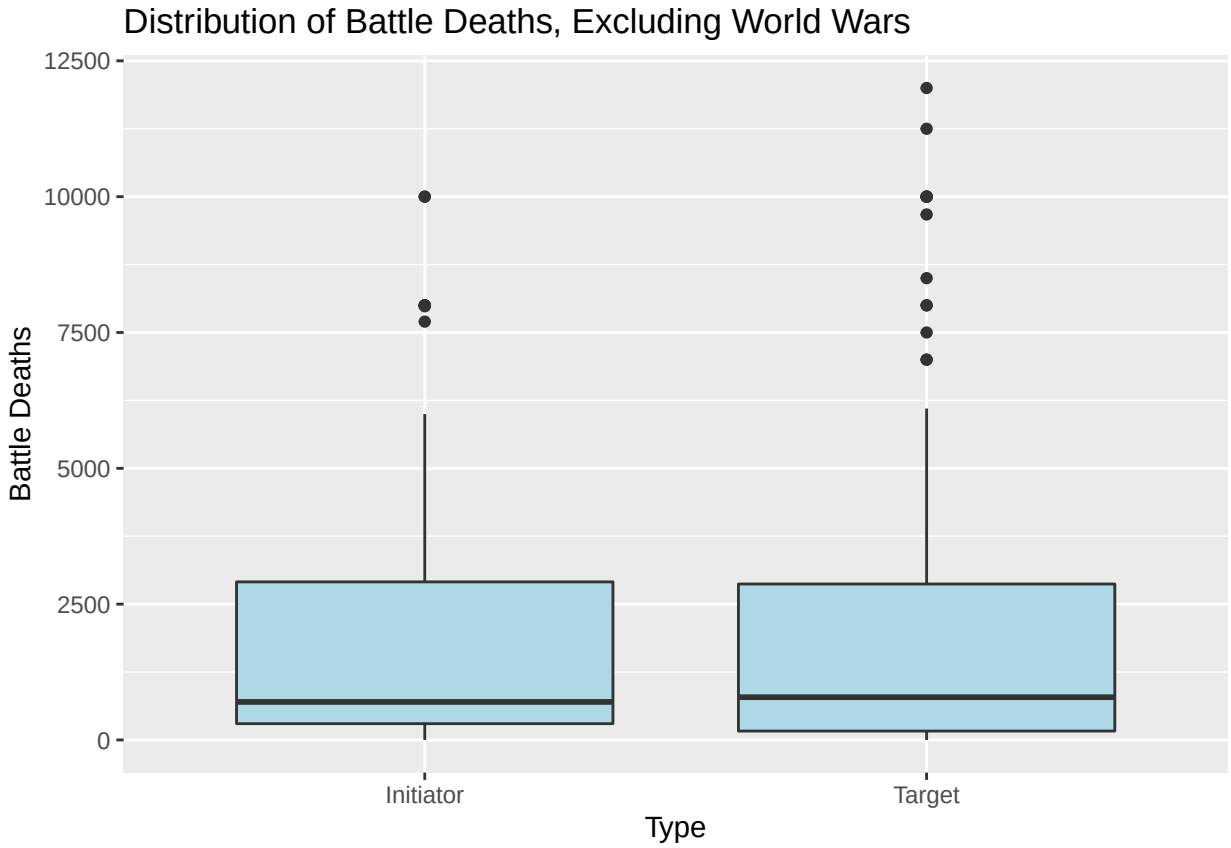


From this plot, we can see that the average number of battle deaths is slightly greater for targets than initiators, but it does not appear to be a big difference. The distributions largely appear to be the same, though there are more targets that are outliers.

The graph below demonstrates that this also appears true if we exclude the World Wars. Overall, it does not appear that initiators generally experience fewer battle deaths than targets.

```
# repeat to see if results hold up when excluding the World Wars
#create a new variable that is missing for any battle deaths greater than 12000
datnww$Death12000 <- ifelse(datnww$`Battle Deaths` <= 12000, datnww$`Battle Deaths`, NA)
#create boxplot of battledeaths less than 12000 by type using ggplot2
ggplot(datnww, aes(x=Initiate, y=Death12000)) + geom_boxplot(fill="light blue") + ggtitle("Distribution
```

```
## Warning: Removed 52 rows containing non-finite values (stat_boxplot).
```



Problem 2

For Problem 2, We will explore the second research question discussed previously regarding the relationship between time and battle deaths. Has the number of battle deaths increased or decreased over time? As previously mentioned, war fighting technology has improved over time, but this technology has reduced the amount of troops needed to conduct a war. Also, technological developments have been made to protect troops. Based on this, I expect that the total number of battle deaths has decreased over time.

While an individual state may suffer 0 battle deaths, for a war to enter the dataset, it must have a minumum of 1,000 battle-related deaths in a 12 month period. This makes the minimum value of war deaths in a war 1,000. Theoretically, the maximum of battle deaths is the total number of combantants involved in the war. I do not have data on the total number of combantants or potential combantants in each war, but presumably the number is too large to be an issue, though it is not infinite.

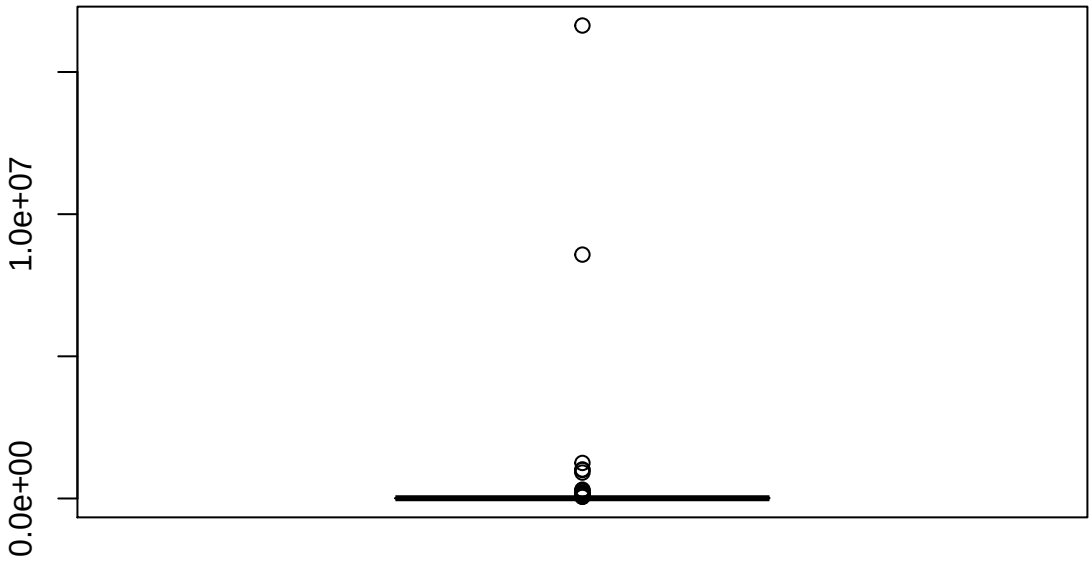
My expectation is that the relationship between battle deaths and time can be modeled as an exponential decay. First we will explore the scatterplot and then we will plot potnetial exponential functions over the data.

```
# create new dataframe
war <- dat2[,c("WarNum","WarName","StartYear1","StartYear2","Battle Deaths", "Deaths1000")]

#create new variable of total deaths and save as dataframe
total <- aggregate(war$`Battle Deaths`, by=list(Category=war$WarNum, Categrory2=war$WarName), FUN=sum ,


#rename column names
colnames(total) <- c("WarNum", "WarName", "TotalDeaths")
```

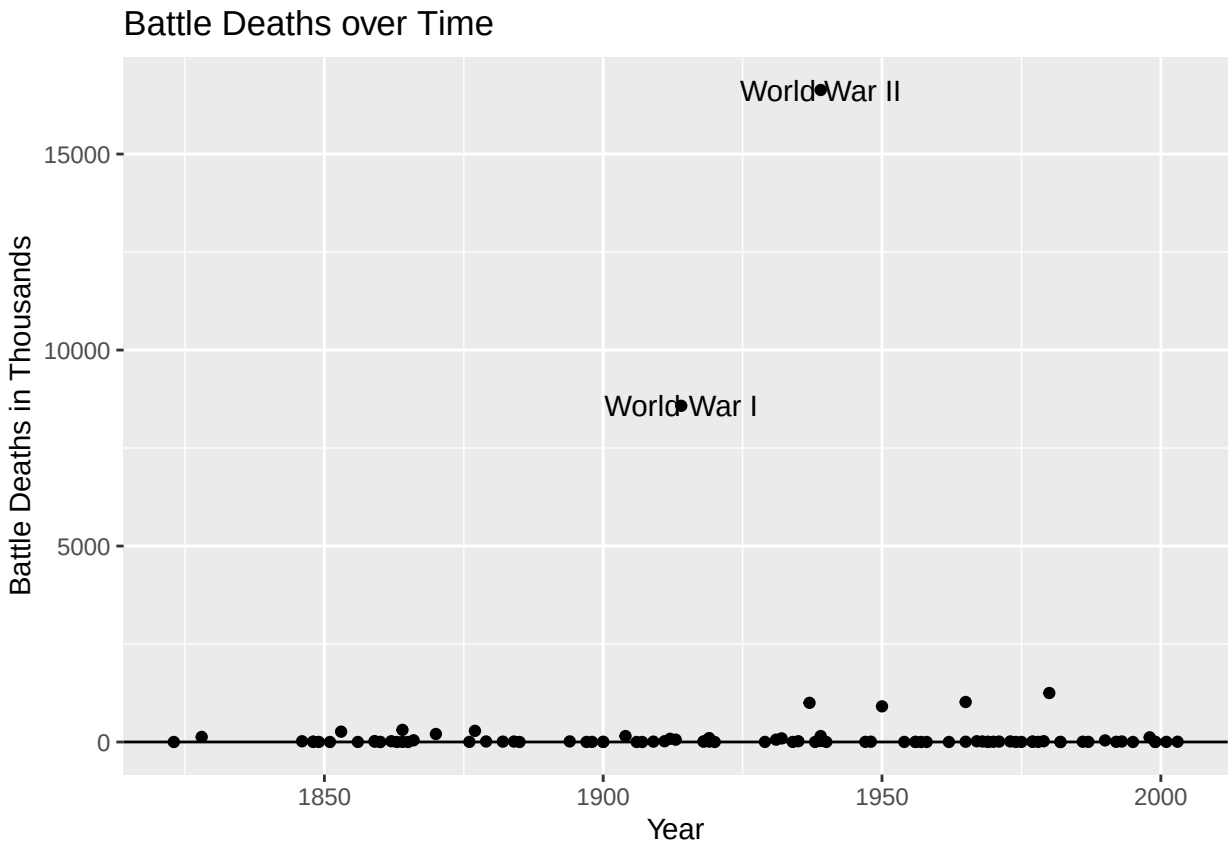
```r
#Look at distribution of total deaths
summary(total$TotalDeaths)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1000     2235     8000   337694    25250 16634907
```

```r
boxplot(total$TotalDeaths)
```



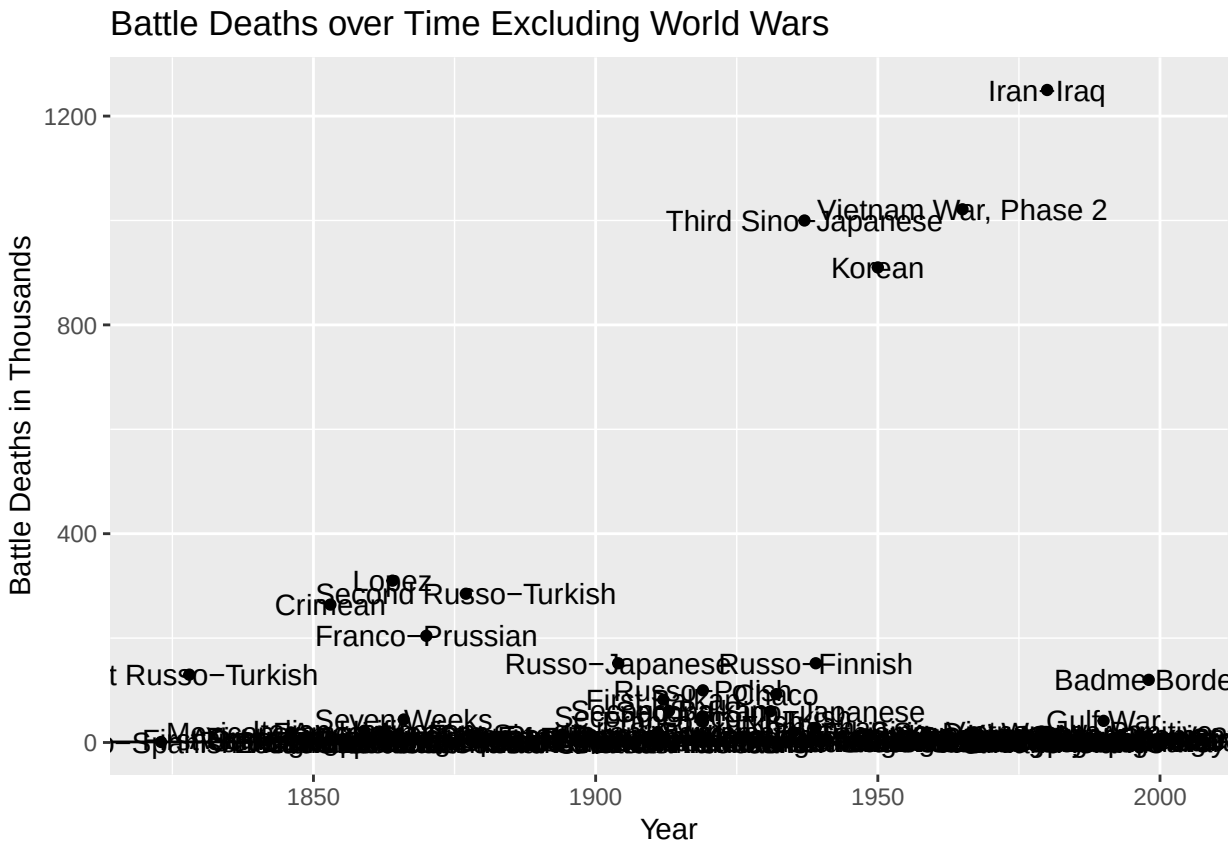

```r
# create new variable
total$Total1000 <- total$TotalDeaths/1000

#create new dataframe to merge with in order to get years
sub <- dat2[,c("WarNum","WarName","StartYear1")]
sub <- sub[order(sub$WarNum, sub$StartYear1),]
test<- sub[!duplicated(sub[,c("WarNum","WarName")]),]

# merge sub dataset and total dataset
total.m <- merge(test,total, by=c("WarNum","WarName"), all=TRUE, sort=TRUE)

#load ggplot2 package
library("ggplot2")

#create scatterplot and label world wars
ggplot(total.m) + aes(StartYear1, Total1000)+geom_point()+ggtitle("Battle Deaths over Time")+ geom_hlin
```



This graph makes it clear that World War I and World War II are outliers in the data. The inclusion of these outliers makes it difficult to see the pattern of the data. Below is a plot without the World Wars.

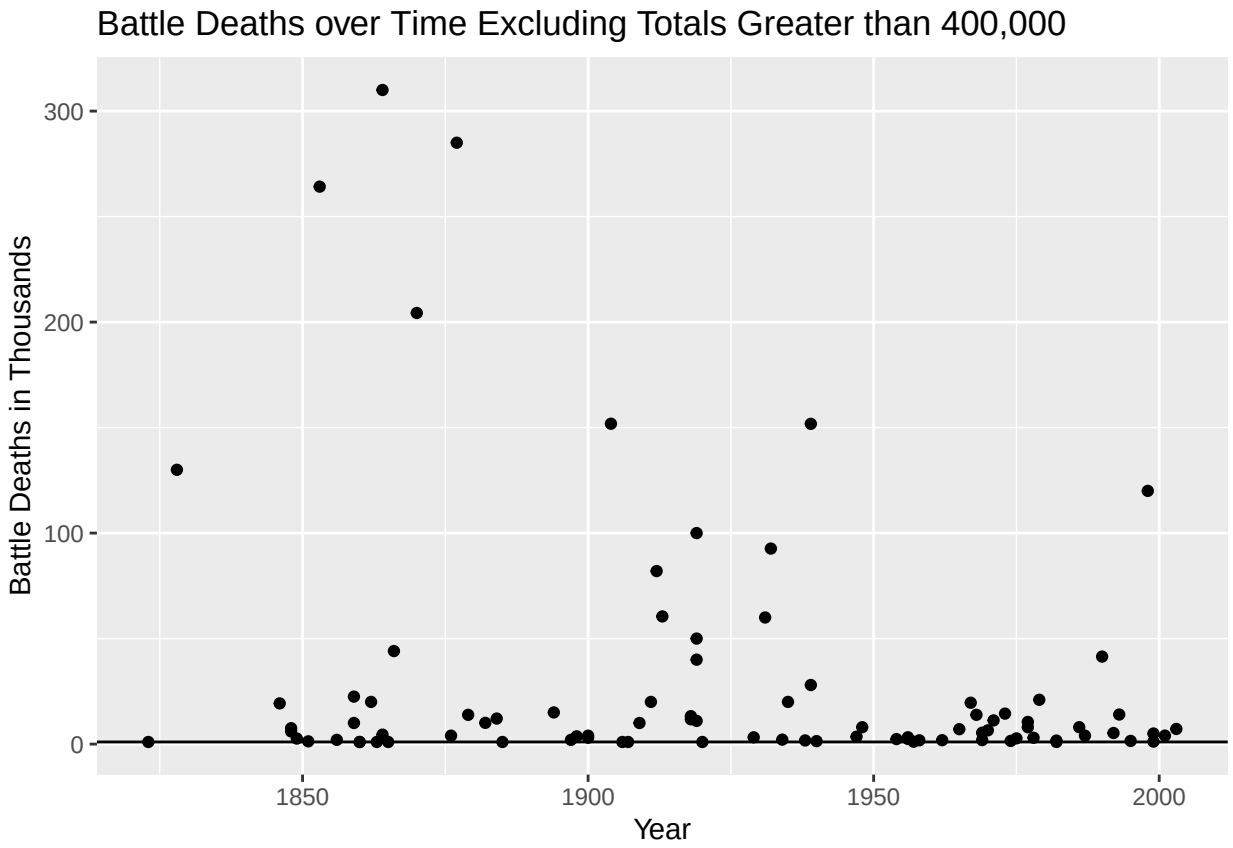
```
# create new dataframe without World War values
totalnww <- total.m[total.m$WarNum != 106,]
totalnww <- totalnww[totalnww$WarNum != 139,]

#create scatterplot without world wars, label so see outliers
ggplot(totalnww) + aes(StartYear1, Total1000)+geom_point()+ggtitle("Battle Deaths over Time Excluding W
```

```
## Warning: Ignoring unknown aesthetics: over_lap
```




```
#create a new variable that is missing for any battle deaths greater than 400,000
total.m$Death400 <- ifelse(total.m$Total1000 <= 400, total.m$Total1000, NA)

#create scatterplot without the large outliers
ggplot(total.m) + aes(StartYear1, Death400)+geom_point()+ggtitle("Battle Deaths over Time Excluding Tota
```

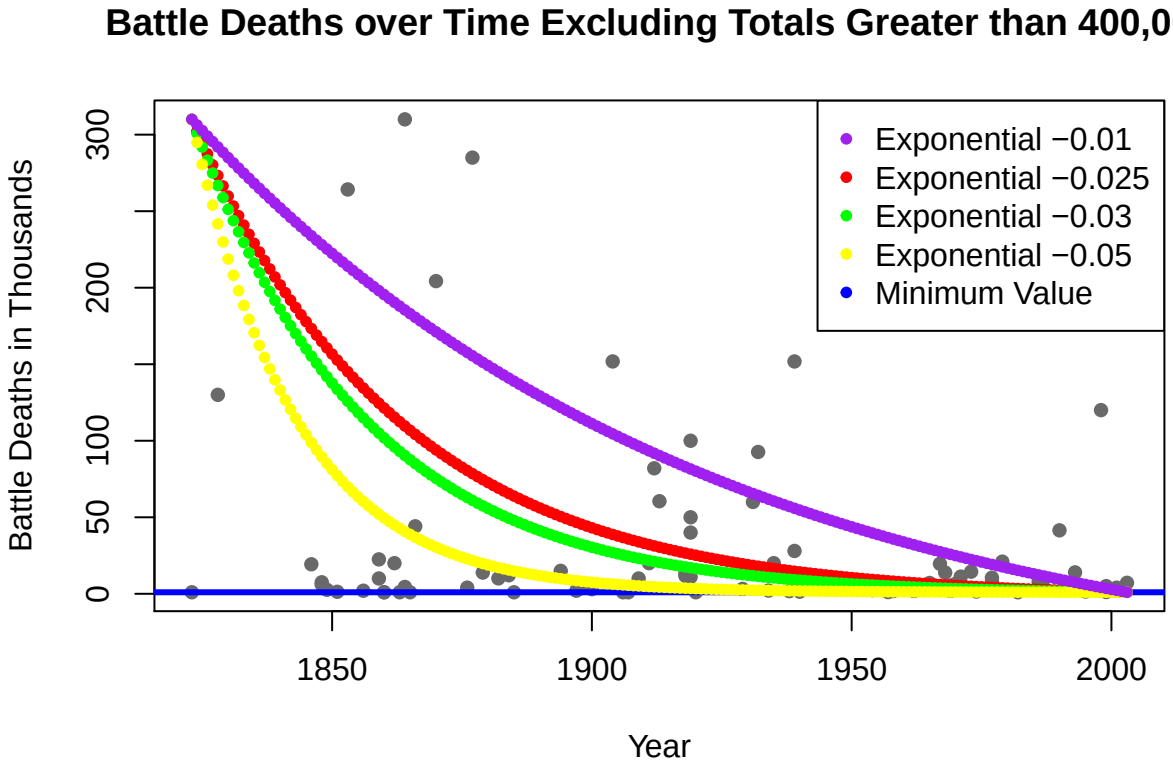
```
## Warning: Removed 6 rows containing missing values (geom_point).
```




```
#Prepare to plot exp function
range(total.m$StartYear1)
```

```
## [1] 1823 2003
```

```
x <- seq(from = 1823, 2003, by =1)
l <- length(x)
index <- seq(from = 1, l, by=1)

#create plot with different exponential lines over the scatterplot
plot(total.m$Death400 ~ total.m$StartYear1, col="dimgray", pch=16, xlab = "Year", ylab="Battle Deaths i
abline(h=1, col="blue", lwd = 3)
title("Battle Deaths over Time Excluding Totals Greater than 400,000")
par(new=TRUE)
plot(exp(-0.025*index)+1, col = "red", axes=FALSE, xlab = "", ylab = "", pch=20)
par(new=TRUE)
plot(exp(-0.03*index)+1, col = "green", axes=FALSE, xlab = "", ylab = "", pch=20)
par(new=TRUE)
plot(exp(-0.05*index)+1, col = "yellow", axes=FALSE, xlab = "", ylab = "", pch=20)
par(new=TRUE)
plot(exp(-0.01*index)+1, col = "purple", axes=FALSE, xlab = "", ylab = "", pch=20)
legend("topright", legend=c("Exponential -0.01","Exponential -0.025","Exponential -0.03", "Exponential
```



After we remove the major outliers, we can see the bulk of the data. The theoretical expectation of an exponential function appears to fit the data fairly well. I think that the red line or the green line appears to fit the data the best. While the theroetical form appears to be exponential, we can run a normal-linear model to determine how well the normal linear model fits the data.

```
#linear regression on data
m_ols <- lm(Total1000 ~ StartYear1, data=total.m)
summary(m_ols)
```

```
##
## Call:
## lm(formula = Total1000 ~ StartYear1, data = total.m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -401.4  -364.5  -312.2  -266.9 16285.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1361.8277  7802.6018  -0.175    0.862
## StartYear1      0.8827     4.0510   0.218    0.828
##
## Residual standard error: 1924 on 93 degrees of freedom
## Multiple R-squared:  0.0005102,  Adjusted R-squared:  -0.01024
## F-statistic: 0.04747 on 1 and 93 DF,  p-value: 0.828
```

```
#linear regression on data excluding World Wars
m2_ols <- lm(Total1000 ~ StartYear1, data=totalnww)
summary(m2_ols)
```

```
##
## Call:
## lm(formula = Total1000 ~ StartYear1, data = totalnww)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -96.31  -80.44  -59.71  -39.94 1158.63
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -544.6264   884.0470  -0.616    0.539
## StartYear1     0.3212     0.4590   0.700    0.486
##
## Residual standard error: 217.8 on 91 degrees of freedom
## Multiple R-squared:  0.005353,   Adjusted R-squared:  -0.005577
## F-statistic: 0.4898 on 1 and 91 DF,  p-value: 0.4858
```

```
#linear regression on data excluding major outliers
m3_ols <- lm(Death400 ~ StartYear1, data=total.m)
summary(m3_ols)
```

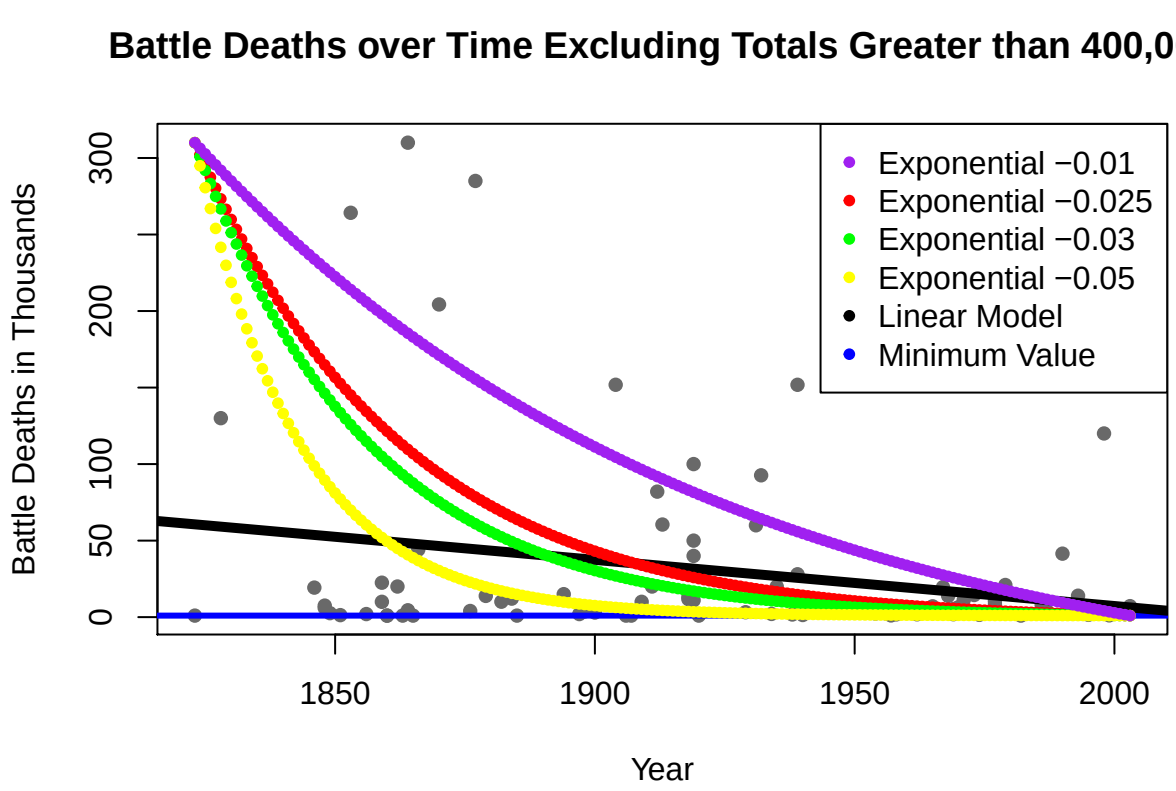
```
##
## Call:
## lm(formula = Death400 ~ StartYear1, data = total.m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.637 -30.384 -14.707  -2.558 261.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 610.4047   245.9241   2.482   0.0150 *
## StartYear1   -0.3016     0.1278  -2.360   0.0205 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.89 on 87 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.06017,    Adjusted R-squared:  0.04937
## F-statistic:  5.57 on 1 and 87 DF,  p-value: 0.0205
```

```
#set coefficients as objects for plotting
m3_ols$coefficients
```

```
## (Intercept)  StartYear1
## 610.4047097  -0.3015731
```

```
int1 <- m3_ols$coefficients[1]
slpe1 <- m3_ols$coefficients[2]


#create plot with different exponential lines over the scatterplot
```

```r
plot(total.m$Death400 ~ total.m$StartYear1, col="dimgray", pch=16, xlab = "Year", ylab="Battle Deaths i
abline(h=1, col="blue", lwd = 3)
abline(a=int1, b=slpe1, col = "black", lty =1, lwd = 5)
title("Battle Deaths over Time Excluding Totals Greater than 400,000")
par(new=TRUE)
plot(exp(-0.025*index)+1, col = "red", axes=FALSE, xlab = "", ylab = "", pch=20)
par(new=TRUE)
plot(exp(-0.03*index)+1, col = "green", axes=FALSE, xlab = "", ylab = "", pch=20)
par(new=TRUE)
plot(exp(-0.05*index)+1, col = "yellow", axes=FALSE, xlab = "", ylab = "", pch=20)
par(new=TRUE)
plot(exp(-0.01*index)+1, col = "purple", axes=FALSE, xlab = "", ylab = "", pch=20)
legend("topright", legend=c("Exponential -0.01","Exponential -0.025","Exponential -0.03", "Exponential
```



The linear model does not appear to capture the data as well as the theoretical exponential functions. It fits the lower values well, but does not capture the larger battle deaths that occured between 1850 and 1950. Linear regression is convenient, but for this data, masks some of the nuances of the data. Thinking about the theoretical expectations and then visualizing the data before running any regressions is helpful in truly exploring the data to answer the question of interest.