```
chechen <- read.csv("chechen.csv")
dhs <- read.csv("dhs_ipv.csv")
```

Question 1

1.1 How many villages were and were not attacked by Russians in the dataset?

In the dataset, 75 villages were attacked and 50 were not.

```
length(unique(chechen$village))
```

```
## [1] 125
```

```
length(unique(chechen$village[chechen$fire == 1]))
```

```
## [1] 75
```

1.2 Create a boxplot of the distribution of deaths occuring in villages in Grozny relative to villages that are not. What do you observe?

There is a higher distribution of deaths occuring in Grozny compared to other villiages; however, the non-Grozny villages have a larger range.

```
groz <- subset(chechen$deaths, chechen$groznyy == 1)
not_groz <- subset(chechen$deaths, chechen$groznyy == 0)

boxplot(groz, not_groz,
main = "distribution of deaths occuring in Grozny vs not",
at = c(1,2),
names = c("Grozny", "Not Grozny"),
las = 2,
col = c("pink","green"),
border = "black",
horizontal = TRUE,
las = .5
)
```

```
## Warning in bxp(list(stats = structure(c(2, 2.5, 3, 4, 5, 0, 0, 0, 1,
## 2), .Dim = c(5L, : Duplicated argument las = 0.5 is disregarded
```

1.3 Using tapply compute the mean number of deaths occuring in villages located in Grozny as well as the mean number of deaths occuring in villages outside of Grozny.

Average number in Grozny is 3.7, and outside is 1.57.

```
grozy <- chechen[chechen$groznyy == 1,]
not_grozy <-chechen[chechen$groznyy == 0,]
avg_g <- tapply(grozy$deaths, grozy$fire == 1, mean)
avg_n <- tapply(not_grozy$deaths, not_grozy$fire == 1, mean)
print(avg_g)
```

```
##    FALSE     TRUE
##       NA 3.714286
```

```
print(avg_n)
```

```
##    FALSE     TRUE
##       NA 1.572368
```

1.4 Compute the difference in the mean number of deaths using the tapply object you created in the prior question.

The difference is 2.14

```
avg_g[2] - avg_n[2]
```

```
##     TRUE
## 2.141917
```

1.5 Now create two dataframes using the subset command: one containing villages in Grozny (Grozny.dat) and tne containing villages not in Grozny (noGrozny.dat). Print the first 6 rows to the screen to confirm that the subsetting worked correctly.

```
noGrozny.dat <- subset(chechen, chechen$groznyy == 0)
Grozny.dat <- subset(chechen, chechen$groznyy == 1)

head(Grozny.dat, 6)
```

```
##                         village groznyy fire deaths preattack postattack
## 37          Grozny (Leninskiy)       1    1      5        21         15
## 38 Staropomyslovskiy (Groznyy)       1    0     NA        16         13
## 57         Grozny  (Zavodskiy)       1    1      3        18         20
## 58      Oktyabr'skiy (Groznyy)       1    0     NA        14         14
## 65 Staropomyslovskiy (Groznyy)       1    0     NA        17         20
## 72          Grozny (Leninskiy)       1    1     10        27         22
```

```
head(noGrozny.dat, 6)
```

```
##         village groznyy fire deaths preattack postattack
## 1   Elistanzhi       0    0     NA         4          3
## 2 Malye Shuani       0    1      0         0          1
## 3     Belgatoy       0    1     34         1          0
## 4 Oktya'brskoe       0    0     NA         0          0
## 5   Chiri-Yurt       0    0     NA         4          5
## 6    Gansolchu       0    1      0         0          0
```

1.6 Did Russian artillery result in a greater number of deaths in Groznyy compared to the villages outside of Groznyy? Conduct the comparison in terms of the mean and median using the new dataframes you created in Question 1.5.

Yes there was more death in Groznyy compared to villages outside. We can see this when comparing the means (3.7 vs 1.6) and the medians (3 to 0).

```
mean(Grozny.dat$deaths[Grozny.dat$fire == 1])
```

```
## [1] 3.714286
```

```
mean(noGrozny.dat$deaths[noGrozny.dat$fire == 1])
```

```
## [1] 1.572368
```

```
median(Grozny.dat$deaths[Grozny.dat$fire == 1])
```

```
## [1] 3
```

```
median(noGrozny.dat$deaths[noGrozny.dat$fire == 1])
```

```
## [1] 0
```

1.7 Working with the original dataframe – chechen – compare the average number of insurgent attacks for villages hit by artillery fire and to the number of insurgent attacks occuring in villages those that were not hit after Russian fire. If this is the only information you had, would you conclude that indiscriminate violence reduces insurgent attacks on the basis of this calculation? Why or why not?

Villages that were fired upon did have fewer insurgent attacks than the others. This suggests that though not by much, that the firing may discourage insurgent attacks after. Though, it is important to note that the quandrant distribution for villiages that were not fired upon is much higher than that of those who experienced the attack which does not support the claim. Ultimately, I would be hesitant to conclude indiscriminate violence reduces insurgent attacks solely on the basis of this calculation.

```
insurgent_fire<- mean(chechen$postattack[chechen$fire == 1])
print(insurgent_fire)
```

```
## [1] 1.496855
```

```
summary(insurgent_fire)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.497   1.497   1.497   1.497   1.497   1.497
```

```
insurgent_no_fire <- mean(chechen$postattack[chechen$fire == 0])
print(insurgent_no_fire)
```

```
## [1] 2.050314
```

```
summary(insurgent_no_fire)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.05    2.05    2.05    2.05    2.05    2.05
```

```
insurgent_fire - insurgent_no_fire
```

```
## [1] -0.5534591
```

1.8 Question 1.7 looks at the difference in means, but lets look at the entire distribution of attacks. Plot the number of attacks experienced by villages that were attacked by the Russians (x-axis) against the number of attacks experienced by villages that were not attacked by the Russians (y-axis). Because villages are either attacked or not, what kind of plot do you need to conduct this comparison? What does the resulting plot show?

The plot shows that between the two experiences, there is not that much differences when comparing the two.

```
p1 = hist(chechen$postattack[chechen$fire == 1], main = "")
```

```
p2 = hist(chechen$postattack[chechen$fire == 0], main = "")
```

```
plot( p1, col=rgb(0,0,1,1/4), xlim=c(0,20), main = "")  # first histogram
plot( p2, col=rgb(1,0,0,1/4), xlim=c(0,20), add=T)  # second
```

```
legend("topright", legend=c("Attacked by Russians", "Not Attacked by Russians"),
       col=c("pink", "green"), lty = 1:2, cex=0.8)
```

1.9 Why do we also want to compare the average number of insurgent attacks in villages that were and were not hit by Russian artillery fire BEFORE the Russian attacks occured? What do you find when you conduct this analysis? What does this mean?

We want to do this in order to verify that the villages post the change were similar prior so we can evaulate the before and after state. We find that the two sets are similar prior to the Russian attacks that occured. We found the average to be similar with only a .037 mean difference. This means we can compare the two.

```
pre_fire <- mean(chechen$preattack[chechen$fire == 1])
print (pre_fire)
```

```
## [1] 2.113208
```

```
pre_not_fire <- mean(chechen$preattack[chechen$fire == 0])
print(pre_not_fire)
```

```
## [1] 2.150943
```

```
pre_fire - pre_not_fire
```

```
## [1] -0.03773585
```

1.10 What kind of plot would allow you to compare how the number of insurgent attacks in a village compares to ohe number of attacks that occur in the same village after Russian artillery fire? Create this plot. Calculate the correlation between attacks before and after Russian fire

The correlation is 0.937

```
p1 = hist(chechen$postattack[chechen$fire == 1], main = "")
```

```
p2 = hist(chechen$preattack[chechen$fire == 1], main = "")
```

```
plot( p1, col=rgb(0,0,1,1/4), xlim=c(0,30), main = "", xlab = "Attacks")  # first histogram
plot( p2, col=rgb(1,0,0,1/4), xlim=c(0,30), add=T)  # second


legend("topright", legend=c("pre", "post"),
       col=c("pink", "purple"), lty = 1:2 )
```

```
cor(chechen$postattack[chechen$fire == 1], chechen$preattack[chechen$fire == 1], method = c("pearson"))
```

```
## [1] 0.9370379
```

1.11 Create a new variable called diffattack by calculating the difference in the number of insurgent attacks before and after the Russian artillery fire. Create a histogram of this difference and interpret. Is this a useful plot for determining whether Russian artillery fire affected the number of insurgent attacks? Why or why Not? If not, can you create a better plot?

This is not a useful plot because it is hard to easily see whether Russian artillery fire affected the number of insurgent attacks. I created a different plot such that one could see the positive vs negative difference more easily. In the new plot, you can see that for the most part, the difference is much negative overall since the trends of the graph are more discernable.

```
chechen$diffattack <- chechen$postattack - chechen$preattack
```

```
hist( chechen$diffattack, col=rgb(0,0,1,1/4), xlim=c(0,4.5), main = "", xlab = "Difference in insugent attacks before/after Russian fire")
```

```
plot(chechen$diffattack,  main = "", type = "l", ylab = "Difference in Insurgent attack", xlab = "", ylim = c(-7,7))
```

1.12 Among the villages shelled by Russians, did the number of insurgent attacks increase after the artillery fires? Give a substantive interpretation of the result.Does this analysis support the claim that indiscriminate violence reduces insurgency attacks? How does this analysis compare to the analysis you conducted in question 1.7? What additional factor does this analysis addresses when compared to the analyses conducted in the previous questions

No, from the villages that were fired upon, the average number attacks decreased by 0.616 when looking at the period before and after. Comparing to the previous analysis, this one supports the suggestion that insurgent attacks decrease after artillery fires.This analysis is a difference in differences. We were looking how the outcomechanges from before the fires compared to after the fires. The factor that this takes into consideration when compared to previous analyses is the assumption that had the villages not been fired upon, we would have examined the same results. This analyses allows us to conclude that the differences are a result of the Russian artillery fire by eliminating confounding variables.

```
mean(chechen$postattack[chechen$fire == 1]) - mean(chechen$preattack[chechen$fire == 1])
```

```
## [1] -0.6163522
```

1.13 Does your conclusion depend on whether the villages are located in Grozny or not? Use the subsets you created in Question 1.5 to examine the effect of indiscriminant violence by region using whatever method of comparison you find most convincing. Justify the choice of comparison.

What do you find and does this analysis make you question the conclusion you reached when answering question 1.12? Why or why not?

I used the difference by difference method in order to compare the two sets and found it to be convincing since the two regions were previously found to be similar enough to compare their differences. When excluding Grozny from the analysis, we can compare the .44 to the -.515 that is examined when including Grozny in the analysis. There difference btween those nubmers is fairly large, so I believe that the conclusion does depend on whether villages are located in Grozny. The region of Grozny is different than the others as found previously, so by only looking at villages not located in Grozny, the conclusion holds.

```
noGrozny.dat$diffattack <- noGrozny.dat$preattack - noGrozny.dat$postattack
mean(noGrozny.dat$diffattack[noGrozny.dat$fire == 1]) - mean(noGrozny.dat$diffattack[noGrozny.dat$fire == 0])
```

```
## [1] 0.4449351
```

```
mean(chechen$diffattack[chechen$fire == 1]) - mean(chechen$diffattack[chechen$fire == 0])
```

```
## [1] -0.5157233
```

Question 2

2.1 Use scatterplots to examine the bivariate relationship between beat_goesout and no_media as well as between beat_goesout and sec_school. Repeat these bivariate graphs between beat_burnfood and no_media, as well as beat_burnfood and sec_school. Be sure to add informative axis labels. Briefly interpret these graphs in light of the hypothesis of the study.

Based on the graphs, there seems to be a relationship in the % of women who believe domestic violence is justified (in situations of burnt food / going out) and both access to media and education. Looking with respects to the hypothesis, there does seem to be a possibliity that women with greater access to media or education are more likely to reject domestic violence.

```
plot(dhs$beat_goesout,dhs$no_media, main = "domestic abuse for going out and media", xlab = "% women who think is justified", ylab = "% wom
```

```
plot(dhs$beat_goesout,dhs$sec_school, main = "domestic abuse for going out and education", xlab = "% women who think is justified", ylab =
```

```
plot(dhs$beat_burnfood,dhs$no_media, main = "domestic abuse for burnt food and media", xlab = "% women who think is justified", ylab = "% w
```

```
plot(dhs$beat_burnfood,dhs$sec_school, main = "domestic abuse for burnt food and education", xlab = "% women who think is justified", ylab
```

2.2 Compute the correlation coefficient between beat_burnfood and media exposure, as well as between beat_burnfood and education. What do these measures tell us about the association between education and media exposure with attitudes towards intimate partner violence?

The first suggests that there is a posivite correlation between media and acceptance of domestic violence whereas the second suggests a negative correlation between education and acceptance of domestic violence. There measures suggests that exposure to media and higher education impact women's attitudes towards intimate partner violence.

```
cor(dhs$beat_burnfood, dhs$no_media, use = c("pairwise.complete.obs"))
```

```
## [1] 0.5967618
```

```
cor(dhs$beat_burnfood, dhs$sec_school, use = c("pairwise.complete.obs"))
```

```
## [1] -0.4760835
```

2.3 We proceed to explore the national-level differences in attitudes towards domestic violence. First, use boxplots to compare the variation in the percentege of beat_burnfood between different regions of the world using region. What are the main differences across regions in terms of the median and dispersion of the distribution? Second, using boxplots examine the distribution of no_media and sec_school by region of the world. Comment on the main differences of the distribution of these variables across regions and what they mean for the results of question 2.3.

There are differences across the regions as some seem to have a higher percentage of agreement with domestic violence over those issues. There seems to be a correlation ebtween a woman's accesss to education and media and their acceptance od domestic violence.

```
x <- dhs$beat_burnfood~dhs$region
boxplot( x, main = "regional agreement with domestic abuse for burnt food", names = c("Asia", "Latin America", "Middle East and Central Asi
```

```
x_2 <- dhs$sec_school~dhs$region
boxplot( x_2, main = "regional agreement with domestic abuse and women's education", names = c("Asia", "Latin America", "Middle East and Ce
```

2.4 An important point of the researcher's hypothesis is that the support towards intimate partner violence should decrease over time, as more women across regions have access to formal schooling and exposure to mass media. To test this idea, using time-series plots, examine the trends in beat_burnfood from 1999-2014 within each region. Label the graph and the trends appropriately. Thinking about the study design, what should we consider before trusting that this plot shows a change over time in attitudes?

We should consider the way in which the data was collected. Looking at the graph, one can see that there are missing gaps in the data for Asia, as well as lookign at the amount of measurements completed by region. For example, when looking at the sumamry we see 84 measurements from SSA and only 19 from the middle east. This could suggest that all of the data recorded was not recorded equally.

```
A <- subset(dhs, dhs$region == "Asia")
year_avg_A <- tapply(A$beat_burnfood, A$year, mean, na.rm = T)

LA <- subset(dhs, dhs$region == "Latin America")
year_avg_LA <- tapply(LA$beat_burnfood, LA$year, mean, na.rm = T)

ME_CA <-subset(dhs, dhs$region == "Middle East and Central Asia")
year_avg_ME_CA <- tapply(ME_CA$beat_burnfood, ME_CA$year, mean, na.rm = T)

SSA <- subset(dhs, dhs$region == "Sub-Saharan Africa")
year_avg_SSA <- tapply(SSA$beat_burnfood, SSA$year, mean, na.rm = T)
```

```
summary(ME_CA)
```

```
##        X          beat_burnfood    beat_goesout     sec_school
##  Min.   :  1.00   Min.   : 0.30   Min.   : 3.10   Min.   :18.50
##  1st Qu.: 24.00   1st Qu.: 4.35   1st Qu.:17.02   1st Qu.:35.30
##  Median : 45.00   Median : 7.35   Median :24.80   Median :43.30
##  Mean   : 56.58   Mean   :12.02   Mean   :26.18   Mean   :46.37
##  3rd Qu.: 72.50   3rd Qu.:13.82   3rd Qu.:36.70   3rd Qu.:54.30
##  Max.   :154.00   Max.   :59.50   Max.   :49.50   Max.   :74.60
##                   NA's   :3       NA's   :3       NA's   :2
##     no_media                 country       year
##  Min.   : 1.500   Egypt          :5   Min.   :2000
##  1st Qu.: 2.800   Jordan         :4   1st Qu.:2003
##  Median : 4.750   Armenia        :3   Median :2006
##  Mean   : 5.844   Albania        :1   Mean   :2006
##  3rd Qu.: 7.200   Azerbaijan     :1   3rd Qu.:2010
##  Max.   :19.100   Kyrgyz Republic:1   Max.   :2014
##  NA's   :1        (Other)        :4
##                          region
##  Asia                        : 0
##  Latin America               : 0
##  Middle East and Central Asia:19
##  Sub-Saharan Africa          : 0
##
##
##
```

```
summary(SSA)
```

```
##        X          beat_burnfood    beat_goesout     sec_school
##  Min.   : 11.00   Min.   : 4.50   Min.   : 5.40   Min.   : 3.10
##  1st Qu.: 49.75   1st Qu.:11.70   1st Qu.:26.30   1st Qu.: 7.60
##  Median : 91.50   Median :18.70   Median :37.70   Median :12.50
##  Mean   : 89.71   Mean   :20.86   Mean   :38.14   Mean   :16.85
##  3rd Qu.:133.25   3rd Qu.:25.65   3rd Qu.:50.50   3rd Qu.:23.00
##  Max.   :160.00   Max.   :64.50   Max.   :82.70   Max.   :69.10
##                   NA's   :17      NA's   :15      NA's   :1
##     no_media                    country        year
##  Min.   : 7.50   Uganda             : 6   Min.   :1999
##  1st Qu.:27.38   Tanzania           : 5   1st Qu.:2004
##  Median :38.75   Nigeria            : 4   Median :2008
##  Mean   :38.66   Senegal            : 4   Mean   :2007
##  3rd Qu.:47.17   Benin              : 3   3rd Qu.:2011
##  Max.   :86.40   Congo (Brazzaville): 3   Max.   :2014
##  NA's   :6       (Other)            :59
##                          region
##  Asia                        : 0
##  Latin America               : 0
##  Middle East and Central Asia: 0
##  Sub-Saharan Africa          :84
##
##
##
```

```
plot(names(year_avg_A), year_avg_A,  xlab = "years", ylab = "% agreeing with domestic violence over burnt food", xlim = c(1999, 2014), col
lines(names(year_avg_LA), year_avg_LA, type = "l", col = "blue")
lines(names(year_avg_ME_CA), year_avg_ME_CA, type = "l", col = "green")
lines(names(year_avg_SSA), year_avg_SSA, type = "l", col = "purple")

# Add a legend to the plot
legend("topright", legend=c("Asia", "Latin America", "Middle East /Central Asia", "Sub-Saharan Africa"),
       col=c("red", "blue", "green", "purple"), lty = 1:2, cex=0.8)
```