Descriptive statistics of variables after preprocessing

The structure of this file is similar to the previous ones: for each variable, we have a small section showing its results with a small analysis or observation of them. Not all variables have been modified, but since we deleted some rows, all of them are affected anyways, so we decided not to exclude any variable from this analysis, even if we’re only showing the small differences in their summary.

#### Variable provstate

As we can see in the barplot, California and New York are the most attacked states, much more than any other state.

#### Variable city

In the following table, we can see the top 10 cities which have the biggest number of attacks recorded in the dataset. We can see that New York City has the most of all the cities by a huge margin, having more than four times the amount of incidents as San Juan, the city in second place. Moreover, taking into account the fact that the dataset records attacks of approximately the last 40 years, only the cities on the top of this table have a big amount of attacks and the rest of them only have occasional incidents: the number quickly goes down.

City name Number of attacks  
439 New York City 449  
566 San Juan 107  
364 Los Angeles 102  
562 San Francisco 97  
682 Washington 80  
399 Miami 69  
128 Chicago 54  
594 Seattle 39  
61 Berkeley 32  
173 Denver 23

#### Variable latitude

Because of the purpose of latitude in this study, it’s not really relevant to analyze and visualize it now.

summary(mydata$latitude)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 17.97 34.10 38.91 36.84 40.70 64.84

#### Variable longitude

Because of the purpose of longitude in this study, it’s not really relevant to analyze and visualize it now.

summary(mydata$longitude)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
-157.86 -117.87 -85.76 -92.13 -73.93 105.27

#### Variable doubtterr

Because of the meaning of this variable, it’s not necessary to give an analysis of it.

summary(mydata$doubtterr)

Mode FALSE TRUE NA's   
logical 2207 398 133

#### Variable success

summary(mydata$success)

Mode FALSE TRUE   
logical 496 2242

As a logical variable we will represent it with a pie chart.

We can see that there are many more successful acts (~82%) than unsuccessful ones (~18%).

#### Variable attacktype1\_txt

Mode(mydata$attacktype1\_txt)

[1] Bombing/Explosion  
8 Levels: Armed Assault Assassination ... Unarmed Assault

The most common method of attack type is “Bombing/Explosion”.

The category which has most acts is “Bombing/Explosion” which has 49% of the acts, followed by “Facility/Infrastructure Attack” with 31%.

#### Variable targtype1\_txt

Mode(mydata$targtype1\_txt)

[1] Business  
21 Levels: Abortion Related Airports & Aircraft ... Violent Political Party

The most common target type is “Business”.

The “Business” target type has 28% of the records.

#### Variable natlty1\_txt

Mode(mydata$natlty1\_txt)

[1] United States  
58 Levels: Angola Argentina Bahamas Bangladesh Brazil Canada ... Yugoslavia

Obviously the vast majority of the attacks are targeted towards the United States since that is the country we are studying.

We will remove the value United States and take the most significative ones in order to make a proper comparison between the other values.

Nationality of victim Number of attacks  
41 Puerto Rico 60  
47 Soviet Union 45  
12 Cuba 20  
26 Israel 14  
34 Mexico 13  
22 International 12  
52 Turkey 12  
23 Iran 11  
45 South Africa 11  
57 Yugoslavia 10  
16 Egypt 9  
20 Haiti 9  
55 Vietnam 9  
54 Venezuela 8  
21 India 6

As we see in those plots there are some nationalities that receive more acts of terrorism than others. Those are the cases of “Puerto Rico” with 60, “Soviet Union” with 45 and “Cuba” with 20.

#### Variable gname

Mode(mydata$gname)

[1] <NA>  
229 Levels: Action Squad African-American extremists ... Zebra killers

Most of the acts of terrorism were either claimed by individuals or an unknown group.

f<-table(mydata$gname)  
f1 = as.data.frame(f)  
f2 = f1[order(-f1$Freq),]  
f2 = subset(f2, f2$Freq > 20)  
colnames(f2) = c ("Group name", "Number of attacks")  
print(f2)

Group name Number of attacks  
12 Anti-Abortion extremists 196  
120 Left-Wing Militants 169  
88 Fuerzas Armadas de Liberacion Nacional (FALN) 109  
143 New World Liberation Front (NWLF) 86  
222 White extremists 85  
52 Black Nationalists 83  
10 Animal Liberation Front (ALF) 76  
197 Student Radicals 71  
109 Jewish Defense League (JDL) 70  
77 Earth Liberation Front (ELF) 66  
146 Omega-7 50  
220 Weather Underground, Weathermen 43  
50 Black Liberation Army 36  
126 Macheteros 36  
18 Anti-Government extremists 33  
57 Chicano Liberation Front 31  
39 Armed Revolutionary Independence Movement (MIRA) 30  
25 Anti-Muslim extremists 27  
112 Jihadi-inspired extremists 27  
116 Ku Klux Klan 25  
53 Black Panthers 24  
165 Puerto Rican Nationalists 23  
196 Strikers 23  
43 Army of God 21  
73 Cuban Exiles 21

As we can see in this barplot there are a lot of group names. The one with more associated acts is “Anti-Abortion extremists” with 196 followed by “Left-Wing Militants” with 169.

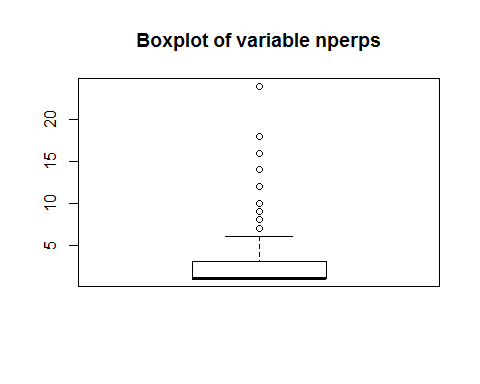
#### Variable nperps

summary(mydata$nperps)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 1.00 1.00 1.00 2.08 3.00 24.00 1941

The minimun value is 1 because it makes no sense that no perpetrator has commited a crime. In preprocessing we corrected those values by changing them to NA.

boxplot(mydata$nperps, main = "Boxplot of variable nperps")



As we can see in this boxplot most of the acts of terrorism in the United States were perpetrated by individuals that worked alone.

skewness(mydata$nperps,na.rm = TRUE)

[1] 4.292201

This value tells us that this variable is asimetric, just as we have seen in the boxplot.

kurtosis(mydata$nperps,na.rm = TRUE)

[1] 32.51557

This value tells us that there is an extreme peak in the distribution where most of the values are clustered. The peak is number 1.

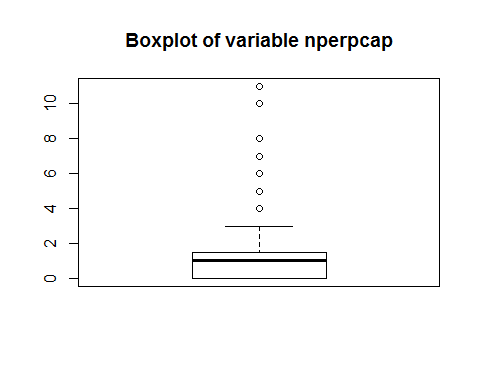
#### Variable nperpcap

summary(mydata$nperpcap)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.000 0.000 1.000 1.239 1.500 11.000 1931

In this case the value 0 makes sense.

boxplot(mydata$nperpcap, main = "Boxplot of variable nperpcap")



As we can see in this boxplot most of the time 0, 1 or 2 perpetrators are captured.

skewness(mydata$nperpcap,na.rm = TRUE)

[1] 2.244441

The positive value of skewness shows us that the variable value distribution is asimetric, its mode being a lower value than the median.

kurtosis(mydata$nperpcap,na.rm = TRUE)

[1] 10.33862

This value tells us that there is a peak in the distribution where most of the values are clustered. The peak values are numbers 0 and 1.

#### Variable claimed

summary(mydata$claimed)

Mode FALSE TRUE NA's   
logical 1194 516 1028

As we can see in this pie chart 70% of the acts were not claimed while 30% of them were.

#### Variable claimmode\_txt

summary(mydata$claimmode\_txt)

Call (post-incident) Call (pre-incident)   
 82 11   
 E-mail Letter   
 9 105   
 Note left at scene Other   
 71 52   
 Personal claim Posted to website, blog, etc.   
 105 32   
 Video NA's   
 2 2269

As we see in the pie chart there are multiple methods of claiming an act with different occurance rates.

#### Variable weaptype1\_txt

summary(mydata$weaptype1\_txt)

Biological Chemical Explosives   
 23 20 1345   
 Fake Weapons Firearms Incendiary   
 4 390 834   
 Melee Other Radiological   
 43 16 1   
Sabotage Equipment Vehicle NA's   
 17 8 37

f<-table(mydata$weaptype1\_txt)  
f1=as.data.frame(f)  
plot\_ly(f1, labels = ~Var1, values = ~Freq, type = 'pie', height = 750) %>%  
 layout(title = 'Pie Chart of variable weaptype1\_txt',  
 xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),  
 yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))

As we see in those plots, there is a big favourite weapon type used during the recorded acts of terrorism which is “Explosives” with 50%, followed by “Incendiary” with 31%.

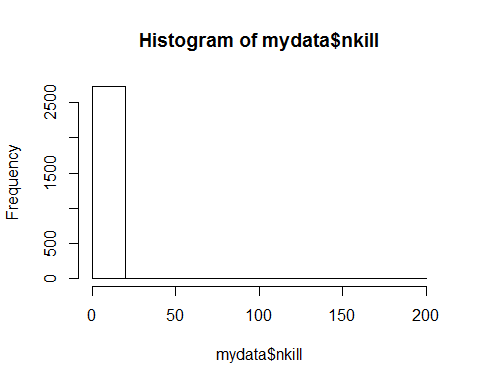
#### Variable nkill

summary(mydata$nkill)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0000 0.0000 0.0000 0.2772 0.0000 190.0000

Even though there are some extreme values like 190, the mean stays really close to 0. This gives us an idea of the amount of 0s this variable contains.

hist(mydata$nkill)



As we can see in this histogram most of the acts of terrorism do not have fatal victims.

skewness(mydata$nkill,na.rm = TRUE)

[1] 42.26456

This value tells us that this variable is widely asimetric, just as we have seen in the histogram

kurtosis(mydata$nkill,na.rm = TRUE)

[1] 1987.84

This value tells us that there is an extreme peak in the distribution where most of the values are clustered. This peak is 0 as we have seen before.

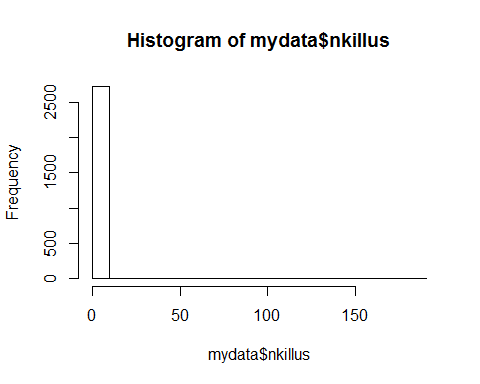
#### Variable nkillus

summary(mydata$nkillus)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0000 0.0000 0.0000 0.2341 0.0000 182.0000

This variable is very similar to nkill.

hist(mydata$nkillus)



As we can see in this histogram most of the acts of terrorism do not have fatal victims of United States citizens.

skewness(mydata$nkillus,na.rm = TRUE)

[1] 43.20675

This value tells us that this variable is widely asimetric, just as we have seen in the histogram.

kurtosis(mydata$nperpcap,na.rm = TRUE)

[1] 10.33862

This value tells us that there is an extreme peak in the distribution where most of the values are clustered. This peak is 0 as we have seen before.

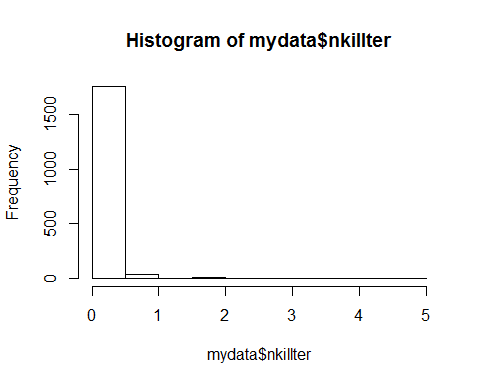
#### Variable nkillter

summary(mydata$nkillter)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.000 0.000 0.000 0.036 0.000 5.000 931

This variable is very focused on 0 considering its max value is 5.

hist(mydata$nkillter)



As we can see in this histogram most of the acts of terrorism do not have fatal victims terrorists.

skewness(mydata$nkillter,na.rm = TRUE)

[1] 10.19833

This positive value shows us that it is an asimetric variable.

kurtosis(mydata$nkillter,na.rm = TRUE)

[1] 140.5823

This positive value shows us that most values are clustered around a peak which is number 0 as we noticed before.

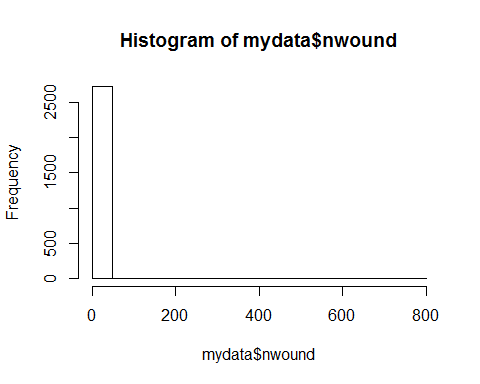
#### Variable nwound

summary(mydata$nwound)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.000 0.000 0.000 1.008 0.000 751.000

The maximum value of 751 is way off the mean. The 1st and 3rd quartile are 0 which tells us that most values are 0.

hist(mydata$nwound)



As we can see in this histogram most of the acts of terrorism do not have wounded victims.

skewness(mydata$nwound,na.rm = TRUE)

[1] 41.2821

This value tells us that this variable is very asimetric, just as we have seen in the histogram.

kurtosis(mydata$nperpcap,na.rm = TRUE)

[1] 10.33862

This value tells us that there is an extreme peak in the distribution where most of the values are clustered. This peak is 0 as we have seen before.

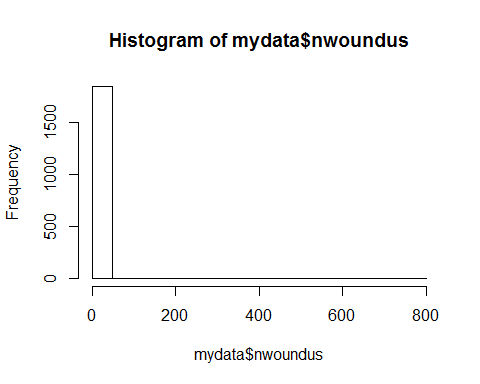
#### Variable nwoundus

summary(mydata$nwoundus)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.00 0.00 0.00 1.07 0.00 751.00 881

The maximum value of 751 is way off the mean. The 1st and 3rd quartile are 0 which tells us that most values are 0.

hist(mydata$nwoundus)



As we can see in this histogram most of the acts of terrorism do not have wounded United States citizens.

skewness(mydata$nwoundus,na.rm = TRUE)

[1] 38.63135

This positive value shows us that the variable is asimetric.

kurtosis(mydata$nperpcap,na.rm = TRUE)

[1] 10.33862

This positive value shows us that the values are clustered around a peak which is 0 in this case.

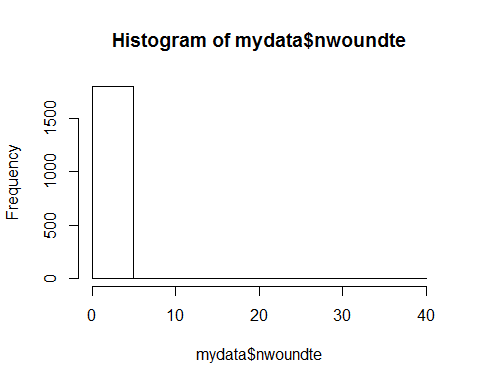
#### Variable nwoundte

summary(mydata$nwoundte)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.0000 0.0000 0.0000 0.0465 0.0000 36.0000 932

This variable is very similar to nkillter. This is related to the fact that the number of perpetrators nperp is 1 most of the times.

hist(mydata$nwoundte)



As we can see in this histogram most of the acts of terrorism do not have wounded terrorists.

skewness(mydata$nwoundte,na.rm = TRUE)

[1] 35.3397

This positive value shows us that this variable is asimetric.

kurtosis(mydata$nwoundte,na.rm = TRUE)

[1] 1364.17

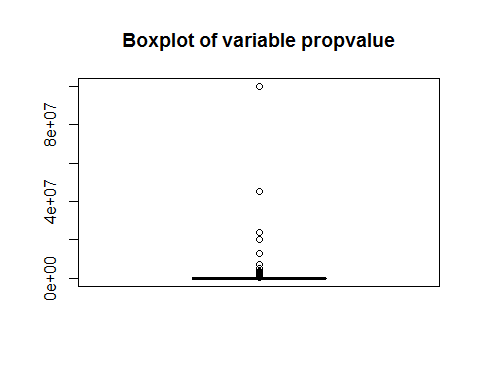
This positive value shows us that there is an extreme peak where all values are clustered around.

#### Variable propvalue

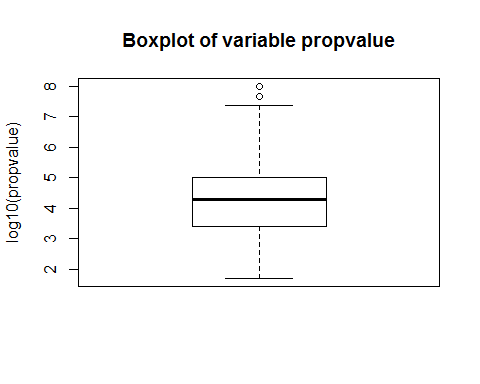
summary(as.numeric(mydata$propvalue))

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0 2500 20000 501607 100000 100000000 2078

mydata.propvaluelog <- sapply(as.vector(mydata$propvalue), function(x) log10(x))  
boxplot(mydata$propvalue, main = "Boxplot of variable propvalue")



boxplot(mydata.propvaluelog, main = "Boxplot of variable propvalue", ylab = "log10(propvalue)")



These are the boxplots for both the variable propvalue and its logarithm, because log(propvalue) is the one that appears to follow a normal distribution. We can observe that the majority of the incidents had very low valued property damage.

skewness(mydata$propvalue[!is.na(mydata$propvalue)])

[1] 18.4448

This value tells us that this variable is very asimetric, just as we have seen before.

kurtosis(mydata$propvalue[!is.na(mydata$propvalue)])

[1] 384.4994

This value shows us there is a big peak distribution which we know happens in the values close to 0.

#### Variable INT\_IDEO

summary(mydata$INT\_IDEO)

Mode FALSE TRUE NA's   
logical 888 434 1416

This pie chart indicates the proportion of ideologically international registered incidents out of the total. Only one third of them, 33%, was considered international by this criteria.

#### Variable INT\_MISC

summary(mydata$INT\_MISC)

Mode FALSE TRUE NA's   
logical 2307 320 111

This pie chart indicates that only 12% of the registered incidents were international.

#### Variable president\_party

summary(mydata$president\_party)

Democratic Republican   
 842 1896

This is the distribution of the number of acts of terrorism and which was the president party at the time of the incident. We can see that the majority of the incidents (69%) happened under a Republican president.

#### Variable state\_governor\_party

summary(mydata$state\_governor\_party)

Democrat   
 1019   
 Democrat, Independent   
 1   
 Democrat, Republican   
 302   
 Democratic-Farmer-Labor   
 8   
 Minnesota Independence Party   
 3   
 New Progressive Party   
 147   
New Progressive Party of Puerto Rico (PNP)   
 1   
 Popular Democratic Party   
 78   
 Republican   
 960   
 NA's   
 219

This is the distribution of the number of acts of terrorism and which was the state governor party at the time of the incident. Under the Democrat and Republican parties, the two main ones in the country, there were a similar amount of incidents (40% and 38% respectively).

#### Variable date

We will not statistically analyse this variable because it makes no sense to check for any distributions or deduce conclusions with the data of the attacks since it is not a continuos variable.