### Preprocess of each variable

Now we start preprocessing the null values of each variable, as well as the incorrect ones and the outliers. We assign the type to the different variables at the end of the script because the incorrect values corrupt the data when assigning the right type.

#### Variable provstate

Thanks to the summary command, we discover that there are 4 rows with NaN values and 1 equal to Unknown, so we start by looking if we can fix them. First, we try to deal with the Nan values by making a subset with them and looking at it.

id provstate city latitude longitude  
769 768 Washington 38.90864 -77.01538  
2199 2198 New York City 40.69713 -73.93135  
2774 2773 Hudson Bend 30.41117 -97.93015  
2780 2779 Manchester 42.99361 -71.48640

As expected, all rows have a null provstate value. However, the city variable isn’t null: this means that we can easily find the state by looking at where these cities are located, using external information if needed, so we can complete the provstate variable with the corresponding values.

mydata$provstate[which(mydata$id == 768)] <- 'District of Columbia'   
mydata$provstate[which(mydata$id == 2198)] <- 'New York'   
mydata$provstate[which(mydata$id == 2773)] <- 'Texas'   
mydata$provstate[which(mydata$id == 2779)] <- 'New Hampshire'

Now we move onto the row with the Unknown value, and study what happened with it.

id provstate city latitude longitude  
1085 1084 Unknown Unknown NA NA

This row has the city value as null, and also lacks its latitude and longitude, so we cannot figure out where this incident took place. Because of that, we decided that it wasn’t worth to keep and chose to delete it.

mydata <- mydata[!mydata$provstate == 'Unknown', ]

After this cleaning process, we shouldn’t have any more problems with the provstate variable, so we can see the result of the preprocess of this variable and check that everything’s alright.

Alabama Alaska Arizona   
 16 1 34   
 Arkansas California Colorado   
 5 595 45   
 Connecticut Delaware District of Columbia   
 17 3 84   
 Florida Georgia Hawaii   
 158 29 4   
 Idaho Illinois Indiana   
 14 112 22   
 Iowa Kansas Kentucky   
 24 14 4   
 Louisiana Maine Maryland   
 22 4 35   
 Massachusetts Michigan Minnesota   
 57 45 25   
 Mississippi Missouri Montana   
 11 39 6   
 Nebraska Nevada New Hampshire   
 24 18 11   
 New Jersey New Mexico New York   
 47 24 516   
 North Carolina North Dakota Ohio   
 34 6 51   
 Oklahoma Oregon Pennsylvania   
 15 67 33   
 Puerto Rico Rhode Island South Carolina   
 248 2 5   
 South Dakota Tennessee Texas   
 9 24 64   
 U.S. Virgin Islands Utah Vermont   
 3 20 5   
 Virginia Washington West Virginia   
 42 101 2   
 Wisconsin Wyoming   
 37 2

As we can see, all the problems have been correctly treated. Now we have to repeat this process for all the other remaining variables: since the steps are the same everytime, we won’t thoroughly explain them again, and instead will highlight the important differences, like the reasoning and process we follow to handle the Nan and Unknown values.

#### Variable city

We see that there are 14 rows without city. We make a subset to study them.

id provstate city latitude longitude  
904 903 Puerto Rico Unknown 18.22083 -66.59015  
905 904 Puerto Rico Unknown 18.22083 -66.59015  
906 905 Puerto Rico Unknown 18.22083 -66.59015  
934 933 Puerto Rico Unknown 18.22083 -66.59015  
935 934 Puerto Rico Unknown 18.22083 -66.59015  
1117 1116 Florida Unknown 27.66483 -81.51575  
1193 1192 Puerto Rico Unknown 18.22083 -66.59015  
1345 1344 Puerto Rico Unknown 18.22083 -66.59015  
1346 1345 Puerto Rico Unknown 18.22083 -66.59015  
1709 1708 Puerto Rico Unknown 18.22083 -66.59015  
1882 1881 New York Unknown 43.00684 -75.04701  
2032 2031 Puerto Rico Unknown 18.22083 -66.59015  
2335 2334 Pennsylvania Unknown 41.20332 -77.19452  
2336 2335 Oregon Unknown 43.80413 -120.55420

These rows have the latitude and longitude variables not empty so we have decided to complete these values using external information as Google Maps, so we can complete the city variable with the corresponding values.

First we put this variable as character for make possible change Unknown values.

mydata$city <- as.character(mydata$city)

Now we can replace missing values of this variable.

mydata$city[which(mydata$provstate == 'Puerto Rico' & mydata$city == 'Unknown')] <- 'Jayuya'  
mydata$city[which(mydata$provstate == 'Florida' & mydata$city == 'Unknown')] <- 'Polk County'  
mydata$city[which(mydata$provstate == 'New York' & mydata$city == 'Unknown')] <- 'Ilion'  
mydata$city[which(mydata$provstate == 'Pennsylvania' & mydata$city == 'Unknown')] <- 'Level Corner'  
mydata$city[which(mydata$provstate == 'Oregon' & mydata$city == 'Unknown')] <- 'Brothers'

#### Variable latitude

summary(mydata$latitude)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 17.97 34.10 38.91 126.77 40.70 40616.00

We see that there are incorrect values, because it impossible that Max. is 40616.00

aux <- mydata[mydata[,"latitude"] > 100,]  
aux <- aux[c('id','provstate','city','latitude','longitude')]  
aux

id provstate city latitude longitude  
234 233 North Carolina Oxford 36312 -78.58820  
239 238 North Carolina Oxford 36312 -78.58820  
250 249 North Carolina Oxford 36312 -78.58820  
251 250 North Carolina Oxford 36312 -78.58820  
252 251 North Carolina Oxford 36312 -78.58820  
1499 1498 New York Inwood 40616 -73.74661  
2396 2395 Arizona Scottsdale 33494 -111.92069

We think these are miss inputs, so we correct them thanks to the city variable.

mydata$latitude[mydata$latitude == 36312] <- 36.312  
mydata$latitude[mydata$latitude == 40616] <- 40.616  
mydata$latitude[mydata$latitude == 33494] <- 33.494

We now can see the result of the preprocess of this variable that corresponds to the US coordinates (including Hawaii and Puerto Rico).

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 17.97 34.10 38.91 36.68 40.70 64.84

#### Variable longitude

summary(mydata$longitude)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
-66266.00 -117.81 -84.51 -115.17 -73.93 105.27

We see that there are incorrect values, because it impossible that Min. is -66266.00

aux <- mydata[mydata[,"longitude"] < -200,]  
aux <- aux[c('id','provstate','city','latitude','longitude')]  
aux

id provstate city latitude longitude  
1904 1903 Puerto Rico Aibonito 18.13996 -66266

We think this is miss input, so we correct them thanks to the city variable.

mydata$longitude[mydata$longitude == -66266] <- -66.266

We now can see the result of the preprocess of this variable that corresponds to the US coordinates (including Hawaii and Puerto Rico).

Min. 1st Qu. Median Mean 3rd Qu. Max.   
-157.86 -117.79 -84.51 -91.82 -73.93 105.27

#### Variable doubtterr

summary(mydata$doubtterr)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 -9.000 0.000 0.000 -0.321 0.000 1.000

As we can see there are null values with the value -9. So we change these values to Nan because we think that writing 0 or making another suposition has no sense in this case.

mydata$doubtterr[mydata$doubtterr == -9] <- NA

#### Variable success

summary(mydata$success)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.000 1.000 1.000 0.824 1.000 1.000

There is nothing to change in this variable, because it has no null values and this is a binary variable, so max and min values are correct.

#### Variable attacktype1\_txt

summary(mydata$attacktype1\_txt)

0 Armed Assault   
 9 281   
 Assassination Bombing/Explosion   
 132 1382   
 Facility/Infrastructure Attack Hijacking   
 862 17   
Hostage Taking (Barricade Incident) Hostage Taking (Kidnapping)   
 62 20   
 Unarmed Assault Unknown   
 59 11

We think that in this case makes sense to use KNN to predict the null values. However, in order to be able to use the KNN method and input the missing values, there has to exist a high correlation between attacktype1\_txt and the numeric variables of the dataset, so first we filter out the non-numeric variables and then compute their correlation with attacktype1\_txt.

mydata.numerics <- Filter(is.numeric, mydata)  
mydata.numerics$attacktype1\_txt <- as.numeric(mydata$attacktype1\_txt)  
cor(mydata.numerics)

After analizing the results we can see that there is no correlation between them and attacktype1\_txt, so we discard the possibility of using KNN. Instead, we replace the NAs with the mode of the variable.

Mode <- function(x) {  
 ux <- unique(x)  
 ux[which.max(tabulate(match(x, ux)))]  
}  
  
moda\_attack\_type <- Mode(mydata$attacktype1\_txt)  
mydata$attacktype1\_txt[which(mydata$attacktype1\_txt == '0' | mydata$attacktype1\_txt == 'Unknown')] <- moda\_attack\_type

Now we can observe the NA’s have disappeared.

Armed Assault Assassination   
 281 132   
 Bombing/Explosion Facility/Infrastructure Attack   
 1402 862   
 Hijacking Hostage Taking (Barricade Incident)   
 17 62   
 Hostage Taking (Kidnapping) Unarmed Assault   
 20 59

#### Variable targtype1\_txt

There are 9 empty values and 11 unknown values, so we substitute them for the attribute mode.

moda\_target\_type <- Mode(mydata$targtype1\_txt)  
mydata$targtype1\_txt[which(mydata$targtype1\_txt == '' | mydata$targtype1\_txt == 'Unknown')] <- moda\_target\_type

We now can see the result of the preprocess of this variable.

Abortion Related Airports & Aircraft   
 253 53   
 Business Educational Institution   
 800 169   
 Food or Water Supply Government (Diplomatic)   
 7 148   
 Government (General) Journalists & Media   
 326 60   
 Maritime Military   
 7 187   
 NGO Other   
 28 2   
 Police Private Citizens & Property   
 170 340   
Religious Figures/Institutions Telecommunication   
 147 10   
 Terrorists/Non-State Militia Tourists   
 8 9   
 Transportation Utilities   
 17 88   
 Violent Political Party   
 6

#### Variable natlty1\_txt

As we can see with the results of summary command, there are incorrect and null values, so first of all we convert all of them to Nan.

mydata$natlty1\_txt[which(mydata$natlty1\_txt == '')] <- NA  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'Tourists: Hernan Diego Mendoza, Diego Enrique Angelini, Alejandro Damian Pagnucco')] <- NA  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'House')] <- NA  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'Church')] <- NA  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'Journalist: Kurt Eichenwald')] <- NA  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'Los Angeles International Airport')] <- NA  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'Students')] <- NA  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'unidentified white man')] <- NA  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'William Long and Quinton I. Ezeagwula, soldiers who were outside of a military recruiting station')] <- NA

Then we also simplify some values.

mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'Multinational')] <- 'International'  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'Virgin Islands (U.S.)')] <- 'United States'  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'West Germany (FRG)')] <- 'Germany'  
mydata$natlty1\_txt[which(mydata$natlty1\_txt == 'Branch')] <- 'United States' # is a county in the U.S. state of Michigan

We now can see the result of our preprocess in this variable.

Angola Argentina   
 1 1   
 Bahamas Bangladesh   
 2 1   
 Brazil Canada   
 1 1   
 Chile China   
 1 4   
 Colombia Costa Rica   
 2 2   
 Croatia Cuba   
 3 24   
 Czechoslovakia Democratic Republic of the Congo   
 1 1   
 Dominican Republic Egypt   
 5 9   
 France Germany   
 2 2   
 Great Britain Haiti   
 3 9   
 India International   
 6 12   
 Iran Iraq   
 11 5   
 Ireland Israel   
 1 14   
 Ivory Coast Japan   
 1 2   
 Jordan Lebanon   
 1 4   
 Liberia Libya   
 1 3   
 Malawi Mexico   
 1 13   
 New Zealand Nicaragua   
 1 2   
 Panama Philippines   
 1 1   
 Poland Portugal   
 1 4   
 Puerto Rico Rhodesia   
 71 1   
 Russia Saudi Arabia   
 2 2   
 South Africa South Korea   
 11 1   
 Soviet Union Spain   
 47 6   
 Switzerland Taiwan   
 5 1   
 Tunisia Turkey   
 1 15   
 United States Uruguay   
 2461 1   
 Venezuela Vietnam   
 8 9   
 West Bank and Gaza Strip Yugoslavia   
 5 11   
 NA's   
 17

#### Variable gname

As we can see there some null values called Unknown. We convert them to Nan because we think that does not make sense to do any prediction or substitution.

mydata$gname[which(mydata$gname == 'Unknown')] <- NA  
mydata$gname[which(mydata$gname == '')] <- NA

We now can see the result of our preprocess.

Anti-Abortion extremists   
 196   
 Left-Wing Militants   
 169   
 Fuerzas Armadas de Liberacion Nacional (FALN)   
 120   
 White extremists   
 87   
 New World Liberation Front (NWLF)   
 86   
 Black Nationalists   
 83   
 Animal Liberation Front (ALF)   
 76   
 Jewish Defense League (JDL)   
 74   
 Student Radicals   
 71   
 Earth Liberation Front (ELF)   
 66   
 Omega-7   
 54   
 Weather Underground, Weathermen   
 45   
 Macheteros   
 37   
 Black Liberation Army   
 36   
 Anti-Government extremists   
 35   
 Chicano Liberation Front   
 31   
 Armed Revolutionary Independence Movement (MIRA)   
 30   
 United Freedom Front (UFF)   
 29   
 Anti-Muslim extremists   
 27   
 Jihadi-inspired extremists   
 27   
 Ku Klux Klan   
 26   
 Puerto Rican Nationalists   
 26   
 Black Panthers   
 24   
 Strikers   
 23   
 Army of God   
 21   
 Cuban Exiles   
 21   
 George Jackson Brigade   
 20   
 May 19 Communist Order   
 20   
 Independent Armed Revolutionary Commandos (CRIA)   
 19   
 Zebra killers   
 19   
 Jewish Armed Resistance   
 17   
 Aryan Republican Army   
 16   
 Organization of Volunteers for the Puerto Rican Revolution   
 15   
 Revolutionary Commandos of the People (CRP)   
 15   
 Anti-Technology extremists   
 14   
 Muslim extremists   
 14   
 The Justice Department   
 14   
 Armed Commandos of Liberation   
 13   
 Croatian Nationalists   
 12   
 Cuban Action   
 12   
 National Front for the Liberation of Cuba (FLNC)   
 12   
 Anti-Semitic extremists   
 10   
 Black September   
 10   
 Chicano Radicals   
 10   
 Neo-Nazi extremists   
 10   
 Luis Boitel Commandos   
 9   
 Coalition to Save the Preserves (CSP)   
 8   
 Jewish Extremists   
 8   
 Justice Commandos for the Armenian Genocide   
 8   
 Secret Cuban Government   
 8   
 The Order (Silent Brotherhood)   
 8   
 Up the IRS, Inc   
 8   
 Anti-Police extremists   
 7   
 Armenian Secret Army for the Liberation of Armenia   
 7   
 Guerrilla Forces for Liberation   
 7   
 Pedro Albizu Campos Revolutionary Forces   
 7   
 Symbionese Liberation Army (SLA)   
 7   
 American Indian Movement   
 6   
 Anti-Government Group   
 6   
 Anti-White extremists   
 6   
 Aryan Nation   
 6   
 Black Muslims   
 6   
 World Church of the Creator   
 6   
 Americans for a Competent Federal Judicial System   
 5   
 Croatian Freedom Fighters   
 5   
 Environmentalists   
 5   
Evan Mecham Eco-Terrorist International Conspiracy (EMETIC)   
 5   
 Fred Hampton Unit of the People's Forces   
 5   
 International Committee Against Nazism   
 5   
 Jamaat-al-Fuqra   
 5   
 Puerto Rican Armed Resistance   
 5   
 Puerto Rican Revolutionary Movement   
 5   
 Red Guerilla Family   
 5   
 Revolutionary Force Seven   
 5   
 Al-Qaida   
 4   
 Anti-Environmentalists   
 4   
 Anti-LGBT extremists   
 4   
 Black Afro Militant Movement   
 4   
 Covenant, Sword and the Arm of the Lord (CSA)   
 4   
 Grupo Estrella   
 4   
 Latin America Anti-Communist Army (LAACA)   
 4   
 National Integration Front (FIN)   
 4   
 New Jewish Defense League   
 4   
 New Year's Gang   
 4   
 Popular Liberation Army (Puerto Rico)   
 4   
Provisional Coordinating Committee for the Defense of Labor   
 4   
 Rajneeshees   
 4   
 Right-wing extremists   
 4   
 Sovereign Citizen   
 4   
 American Servicemen's Union (ASU)   
 3   
 Anarchists   
 3   
 Animal Rights extremists   
 3   
 Anti-Castro Group   
 3   
 Anti-Gun Control extremists   
 3   
 Armed Forces of Popular Resistance (FARP)   
 3   
 Black Revolutionary Assault Team   
 3   
 Earth First!   
 3   
 Front for the National Liberation of Puerto Rico   
 3   
 (Other)   
 193   
 NA's   
 580

#### Variable nperps

summary(as.factor(mydata$nperps))

-99 0 1 2 3 4 5 6 7 8 9 10 12 14 16   
1028 10 464 133 96 65 25 8 5 5 1 4 4 1 1   
 18 24 200 400 NA's   
 1 1 1 1 981

We can see that there are some missing values like -99 and some outliers like 200, 400. Also we think that value 0 maybe is a miss input of this variable, because we can’t have a terrorist attack without perpetrators, so we consider this values as unknown.

We will change the null values to NA and delete the outliers.

mydata$nperps[mydata$nperps == -99] <- NA  
mydata$nperps[mydata$nperps == 0 ] <- NA  
mydata$nperps[mydata$nperps == 200] <- NA  
mydata$nperps[mydata$nperps == 400] <- NA

We now can see the result of the preprocess of this variable.

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 1.000 1.000 1.000 2.082 3.000 24.000 2021

#### Variable nperpcap

summary(as.factor(mydata$nperpcap))

-99 0 1 2 3 4 5 6 7 8 10 11 NA's   
 963 278 337 78 54 43 16 3 2 5 2 1 1053

We can see that there are a lot of -99 that we will need to convert to NA for consistency.

mydata$nperpcap[mydata$nperpcap == -99] <- NA

We now can see the result of the preprocess of this variable.

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.000 0.000 1.000 1.233 1.000 11.000 2016

#### Variable claimed

summary(as.factor(mydata$claimed))

-99 -9 0 1 3 4 NA's   
 1 43 1221 518 2 1 1049

We can see that there are some missing values like -9 and -99. There are also strange values like 3 and 4 that make no sense. We will change all of them to NA.

mydata$claimed[mydata$claimed == -9] <- NA  
mydata$claimed[mydata$claimed == -99] <- NA  
mydata$claimed[mydata$claimed == 3] <- NA  
mydata$claimed[mydata$claimed == 4] <- NA

We now can see the result of the preprocess of this variable.

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.0000 0.0000 0.0000 0.2979 1.0000 1.0000 1096

####. Variable claimmode\_txt

We can see that there are some values that make no sense like 0 and 1. We also have some missing values as Unknown and ’’. We will now change all these values to NA.

mydata$claimmode\_txt[mydata$claimmode\_txt == '0'] <- NA  
mydata$claimmode\_txt[mydata$claimmode\_txt == '1'] <- NA  
mydata$claimmode\_txt[mydata$claimmode\_txt == 'Unknown'] <- NA  
mydata$claimmode\_txt[mydata$claimmode\_txt == ''] <- NA

We now can see the result of the preprocess of this variable.

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   
 4 2364

#### Variable weaptype1\_txt

We can see that the value 0 makes no sense. We also have some missing values as ’’. We will now change all these values to NA.

mydata$weaptype1\_txt[mydata$weaptype1\_txt == '0'] <- NA  
mydata$weaptype1\_txt[mydata$weaptype1\_txt == ''] <- NA  
mydata$weaptype1\_txt[mydata$weaptype1\_txt == 'Unknown'] <- NA

We also will simplify one factor.

mydata$weaptype1\_txt <- as.character(mydata$weaptype1\_txt)  
mydata$weaptype1\_txt[mydata$weaptype1\_txt == 'Vehicle (not to include vehicle-borne explosives, i.e., car or truck bombs)'] <- 'Vehicle'  
mydata$weaptype1\_txt <- as.factor(mydata$weaptype1\_txt)

We now can see the result of the preprocess of this variable.

Biological Chemical Explosives   
 24 21 1402   
 Fake Weapons Firearms Incendiary   
 4 400 844   
 Melee Other Radiological   
 43 16 1   
Sabotage Equipment Vehicle NA's   
 18 10 52

#### Variable nkill

After looking on the results of the summary command we can see that variable has some problems as outliers and miss inputs.

We first analyze the two major numbers: 1383 and 1384, our outliers.

id provstate city latitude longitude date  
2416 2415 New York New York City 40.69713 -73.93135 9-11-2001

id provstate city latitude longitude date  
2415 2414 New York New York City 40.69713 -73.93135 9-11-2001

Thanks to the date we assume that these attacks correspond to the 9/11. We consider these values as outliers because we want to analyse the usual terrorist attacks. So we will delete these rows.

mydata <- mydata[!(mydata$nkill==1383 | mydata$nkill==1384),]

Also we can see that there are a lot wrongly introduced values, we will change them all to NA.

mydata$nkill[mydata$nkill == ''] <- NA  
mydata$nkill[mydata$nkill == 'A knife and a vehicle were used in the attack.' ] <- NA  
mydata$nkill[mydata$nkill == 'A rental pickup truck and replica firearms were used in the attack.' ] <- NA  
mydata$nkill[mydata$nkill == 'A strobe light GIF sent via Twitter was used in the attack.' ] <- NA  
mydata$nkill[mydata$nkill == 'An SKS semi-automatic rifle' ] <- NA  
mydata$nkill[mydata$nkill == 'four containers of a diesel and gasoline mixture, placed at two locations in the office, ignited via timed ignition devices.' ] <- NA  
mydata$nkill[mydata$nkill == 'Gasoline was used in the attack.' ] <- NA  
mydata$nkill[mydata$nkill == 'incendiaries' ] <- NA  
mydata$nkill[mydata$nkill == 'knives, machetes, meat cleavers, metal cutters' ] <- NA  
mydata$nkill[mydata$nkill == 'ruger .22 caliber semi-automatic pistol' ] <- NA  
mydata$nkill[mydata$nkill == 'Shotgun; revolver; pipe bomb' ] <- NA

Since the number of null values is small with respect to the total number of attacks, we will change these values by the median.

mydata$nkill <- as.character(mydata$nkill) # avoid unforeseen results  
mydata$nkill <- as.numeric(mydata$nkill)  
nkill\_without\_nulls <- subset(mydata, !is.na(nkill))  
median\_nkill <- median(nkill\_without\_nulls$nkill)  
mydata$nkill[which(is.na(mydata$nkill))] <- median\_nkill

We now can see the result of the preprocess of this variable.

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0000 0.0000 0.0000 0.3491 0.0000 190.0000

####. Variable nkillus

summary(mydata$nkillus)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.0000 0.0000 0.0000 0.4309 0.0000 182.0000 953

As before, we substitute the nan values with the median.

nkillus\_without\_nulls <- subset(mydata, !is.na(nkillus))  
median\_nkillus <- median(nkillus\_without\_nulls$nkillus)  
mydata$nkillus[which(is.na(mydata$nkillus))] <- median\_nkillus

We can now see the result of the preprocess of this variable.

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0000 0.0000 0.0000 0.2859 0.0000 182.0000

#### Variable nkillter

summary(mydata$nkillter)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.0000 0.0000 0.0000 0.0361 0.0000 5.0000 1003

There is nothing to change in this variable because the number of null values is too huge and we don’t want to risc more by replace then with median value.

#### Variable nwound

summary(mydata$nwound)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.000 0.000 0.000 1.555 0.000 851.000 93

In this variable we decided not to change the NA values, because we do not have the reliable criteria to replace this values.

But after analyzing the summary, we were able to detect some outliers, so we will delete the outliers rows.

mydata <- mydata[!(mydata$nwound==650 | mydata$nwound==751 | mydata$nwound==851),]

We can now see the result of the preprocess of this variable.

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.0000 0.0000 0.0000 0.7337 0.0000 151.0000 93

#### Variable nwoundus

summary(mydata$nwoundus)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.0000 0.0000 0.0000 0.6659 0.0000 151.0000 974

There is nothing to do with the number of null values because there a lot of them. Then we decided to leave this variable as it is now, and to treat the NA’s later in the case that is necessary.

#### 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 1025

In this case it is the same as in the previous variable, so there is nothing to change in this variable because there are many null values.

#### Variable propvalue

We can see with summary, that there are some values that make no sense such as the different sentences. We also have some missing values as -99 and ’’. We will now change all these values to NA.

mydata$propvalue[mydata$propvalue == '-99'] <- NA  
mydata$propvalue[mydata$propvalue == ''] <- NA  
mydata$propvalue[mydata$propvalue == 'Minor (likely < $1 million)'] <- NA  
mydata$propvalue[mydata$propvalue == 'Major (likely >= $1 million but < $1 billion)'] <- NA

We can now see the result of the preprocess of this variable. We will not substitue with the null values because there a lot of them about the total number of attacks.

1000 5000 100000 2000 10000 3000 50000   
 42 37 35 33 27 25 22   
 2500 20000 25000 75000 500000 200000 30000   
 19 18 18 18 17 15 15   
 250000 500 15000 1500 40000 0 1000000   
 12 12 11 9 9 8 8   
 300000 60000 1500000 400 4000 70000 150000   
 8 7 6 6 6 6 5   
 200 90000 100 17000 2000000 2500000 35000   
 5 5 4 4 4 4 4   
 6000 8000 300 333 333333 400000 50   
 4 4 3 3 3 3 3   
 600000 700 7000 700000 80000 9000 105000   
 3 3 3 3 3 3 2   
 11000 12000 125000 135000 150 1600 175000   
 2 2 2 2 2 2 2   
 18000 250 26000 29000 3000000 350000 450000   
 2 2 2 2 2 2 2   
 5000000 5500 65000 7500 82000 100000000 1041696   
 2 2 2 2 2 1 1   
 110000 118000 11890 1200000 1211388 122000 126000   
 1 1 1 1 1 1 1   
 1300 13000 13000000 135 13500 13800 153000   
 1 1 1 1 1 1 1   
 162500 16606 1689593 1750 1800 180000 1830000   
 1 1 1 1 1 1 1   
 187500 1900 20000000 207497 210000 21500 2200   
 1 1 1 1 1 1 1   
 (Other) NA's   
 65 2170

#### Variable INT\_IDEO

We can see that there are some values that make no sense such as the different sentences. We also have some missing values as -9. We will now change all these values to NA.

mydata$INT\_IDEO[mydata$INT\_IDEO == '-9'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == ' revolutionary cells says it left the device at Shaklee in Pleasanton and threatens more violence,\\" Contra Costa Times October, 1 2003."'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'Anti-Abortion Project 2010'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'Brent L. Smith and Kelly R. Damphousse, \\ Patterns of precursor behaviors in the life span of a U.S. environmental terrorist group,\\" Criminology & Public Policy, Volume 8, Issue 3, 2009."'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'CETIS'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'Eco Project 2010'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'FBI, \\Terrorism in the United States: 1999,\\" Counterterrosism Threat Assessment and Warning Unit, Counterterrorism Division, FBI, DOJ, 1998."'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'FBI, \\Terrorism in the United States: 1999,\\" Counterterrosism Threat Assessment and Warning Unit, Counterterrorism Division, FBI, DOJ, 1999."'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'Hewitt Project'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'START Primary Collection'] <- NA  
mydata$INT\_IDEO[mydata$INT\_IDEO == 'UMD Miscellaneous'] <- NA

We can now see the result of the preprocess of this variable.

0 1 NA's   
 888 433 1509

#### Variable INT\_MISC

We can see that there are some values that make no sense such as the different sentences. We also have some missing values as -9. We will now change all these values to NA.

mydata$INT\_MISC[mydata$INT\_MISC == '-9'] <- NA  
mydata$INT\_MISC[mydata$INT\_MISC == 'CETIS'] <- NA  
mydata$INT\_MISC[mydata$INT\_MISC == 'Eco Project 2010'] <- NA  
mydata$INT\_MISC[mydata$INT\_MISC == 'Eric Collins, \\Animal Rights Group Says It Started Fire at Oregon Meat Plant,\\" The Register Guard, June 1, 1999."'] <- NA  
mydata$INT\_MISC[mydata$INT\_MISC == 'Legal Affairs Editors, \\Animal Liberation Front and Earth Liberation Front Members Sentenced in Oregon for Acts of Eco-Terrorism in Five Western States,\\" PR Newswire Association LLC, Public Interest Services, June 5, 2007."'] <- NA

We can now see the result of the preprocess of this variable.

0 1 NA's   
2306 320 204

#### Variable president\_party

summary(mydata$president\_party)

Democratic Republican NA's   
 842 1895 93

In this case it is a variable that we add to the dataset, so we consider there is nothing to do with this variable.

#### Variable state\_governor\_party

We can see that there are some values that are typos like the Democratic instead of Democrat or Domecratic-Farmer-Labor instead of Democratic-Farmer-Labor. We also have some missing values as Nan that we will change into NA.

mydata$state\_governor\_party[mydata$state\_governor\_party == ''] <- NA  
mydata$state\_governor\_party[mydata$state\_governor\_party == 'Nan'] <- NA  
mydata$state\_governor\_party[mydata$state\_governor\_party == 'Democratic'] <- 'Democrat'  
mydata$state\_governor\_party[mydata$state\_governor\_party == 'Domecratic-Farmer-Labor'] <- 'Democratic-Farmer-Labor'

We can now see the result of the preprocess of this variable.

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   
 959   
 NA's   
 312

#### Variable date

We create a sorted subset to see clearly if all the date values are correct.

dateSubset <- mydata[c('id','date')]  
dateSubset[order(dateSubset$date),]

As we can observe that there are some values that have an incorrect value in the day attribute, a zero. We convert these values to 1 because we are not taking into account this characteristic in our study: we will study the year attribute, not the day of the attack.

mydata$date <- as.character(mydata$date)  
mydata$date[which(mydata$date == '1-0-1981')] <- '1-1-1981'  
mydata$date[which(mydata$date == '1-0-1989')] <- '1-1-1989'  
mydata$date[which(mydata$date == '1-0-1994')] <- '1-1-1994'  
mydata$date[which(mydata$date == '1-0-1995')] <- '1-1-1995'  
mydata$date[which(mydata$date == '2-0-1978')] <- '2-1-1978'  
mydata$date[which(mydata$date == '3-0-1972')] <- '3-1-1972'  
mydata$date[which(mydata$date == '3-0-1974')] <- '3-1-1974'  
mydata$date[which(mydata$date == '5-0-1970')] <- '5-1-1970'  
mydata$date[which(mydata$date == '5-0-1980')] <- '5-1-1980'  
mydata$date[which(mydata$date == '5-0-1999')] <- '5-1-1999'  
mydata$date[which(mydata$date == '6-0-1998')] <- '6-1-1998'  
mydata$date[which(mydata$date == '7-0-1970')] <- '7-1-1970'  
mydata$date[which(mydata$date == '7-0-1971')] <- '7-1-1971'  
mydata$date[which(mydata$date == '8-0-1987')] <- '8-1-1987'  
mydata$date[which(mydata$date == '8-0-1999')] <- '8-1-1999'  
mydata$date[which(mydata$date == '9-0-1981')] <- '9-1-1981'  
mydata$date[which(mydata$date == '9-0-2005')] <- '9-1-2005'  
mydata$date[which(mydata$date == '10-0-1981')] <- '10-1-1981'  
mydata$date[which(mydata$date == '10-0-2010')] <- '10-1-2010'  
mydata$date[which(mydata$date == '11-0-1977')] <- '11-1-1977'  
mydata$date[which(mydata$date == '11-0-1990')] <- '11-1-1990'  
mydata$date[which(mydata$date == '11-0-2010')] <- '11-1-2010'  
mydata$date[which(mydata$date == '12-0-1981')] <- '12-1-1981'  
mydata$date[which(mydata$date == '12-0-1992')] <- '12-1-1992'

### Delete rows with lots of null values

After doing all these preprocessing steps for each variable, we have decided that we might be interested in deleting the rows that have a number of nulls per columns that exceed a limit that we set. For this we would do the following steps, but once trying to establish different limits we saw that the dataset was reduced considerably, therefore we have decided to neutralize this step by putting a limit equal to the number of columns in the way that we keep all the rows of the dataset.

We have done that this way, because it is always possible to delete things later, but it would be difficult to recover them if necessary.

So first, we count the number of nulls in each row.

na\_count <- apply(mydata, 1, function(x) sum(is.na(x)))  
summary(na\_count)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.000 3.000 5.000 6.033 8.000 28.000

We delete all rows with all the columns equal to Nan.

mydata$na\_count = na\_count  
mydata<-mydata[!(mydata$na\_count>27),]

We can now see the result of our preprocess. The final dataset have the following rows:

[1] 2737

### Assigning the right type to each variable

Final pass of our preprocessing is to change the types to the right ones.

mydata$id <- as.numeric(mydata$id)  
mydata$provstate <- as.factor(mydata$provstate)  
mydata$city <- as.factor(mydata$city)  
mydata$latitude <- as.numeric(mydata$latitude)  
mydata$longitude <- as.numeric(mydata$longitude)  
mydata$doubtterr <- as.numeric(mydata$doubtterr) # avoid unforeseen results  
mydata$doubtterr <- as.logical(mydata$doubtterr)  
mydata$success <- as.numeric(mydata$success) # avoid unforeseen results  
mydata$success <- as.logical(mydata$success)  
mydata$attacktype1\_txt <- as.factor(mydata$attacktype1\_txt)  
mydata$targtype1\_txt <- as.factor(mydata$targtype1\_txt)  
mydata$natlty1\_txt <- as.factor(mydata$natlty1\_txt)  
mydata$gname <- as.factor(mydata$gname)  
mydata$nperps <- as.numeric(mydata$nperps)  
mydata$nperpcap <- as.numeric(mydata$nperpcap)  
mydata$claimed <- as.numeric(mydata$claimed) # avoid unforeseen results  
mydata$claimed <- as.logical(mydata$claimed)  
mydata$claimmode\_txt <- as.factor(mydata$claimmode\_txt)  
mydata$weaptype1\_txt <- as.factor(mydata$weaptype1\_txt)  
mydata$nkill <- as.character(mydata$nkill) # avoid unforeseen results  
mydata$nkill <- as.numeric(mydata$nkill)  
mydata$nkillus <- as.numeric(mydata$nkillus)  
mydata$nkillter <- as.numeric(mydata$nkillter)  
mydata$nwound <- as.numeric(mydata$nwound)  
mydata$nwoundus <- as.numeric(mydata$nwoundus)  
mydata$nwoundte <- as.numeric(mydata$nwoundte)  
mydata$propvalue <- as.character(mydata$propvalue) # avoid unforeseen results  
mydata$propvalue <- as.numeric(mydata$propvalue)   
mydata$INT\_IDEO <- as.character(mydata$INT\_IDEO);mydata$INT\_IDEO <- as.numeric(mydata$INT\_IDEO) # avoid unforeseen results  
mydata$INT\_IDEO <- as.logical(mydata$INT\_IDEO)  
mydata$INT\_MISC <- as.character(mydata$INT\_MISC); mydata$INT\_MISC <- as.numeric(mydata$INT\_MISC) # avoid unforeseen results  
mydata$INT\_MISC <- as.logical(mydata$INT\_MISC)  
mydata$president\_party <- as.factor(mydata$president\_party)  
mydata$state\_governor\_party <- as.factor(mydata$state\_governor\_party)  
mydata$date <- as.Date(mydata$date,format("%m-%d-%Y"))

We can now check that all variables have the right type assigned.

id provstate city   
 "numeric" "factor" "factor"   
 latitude longitude doubtterr   
 "numeric" "numeric" "logical"   
 success attacktype1\_txt targtype1\_txt   
 "logical" "factor" "factor"   
 natlty1\_txt gname nperps   
 "factor" "factor" "numeric"   
 nperpcap claimed claimmode\_txt   
 "numeric" "logical" "factor"   
 weaptype1\_txt nkill nkillus   
 "factor" "numeric" "numeric"   
 nkillter nwound nwoundus   
 "numeric" "numeric" "numeric"   
 nwoundte propvalue INT\_IDEO   
 "numeric" "numeric" "logical"   
 INT\_MISC president\_party state\_governor\_party   
 "logical" "factor" "factor"   
 date   
 "Date"

### Saving the preprocessed dataset

Finally we save our preprocessed dataset in a csv.

write.csv(mydata, file = "preprocessed\_dataset.csv")