R Notebook

Where is this data from?

It is from the ACLED database. It is a wellknown project and it is hugh. There are some data scientists, social schientists and professors studying on project.

Loading packages.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------- tidyve
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------------------------------------------- tidyverse_co
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)  #Used for data manipulation of the data set
library(ggplot2) #Used for plotting
library(readr) #Used for reading files
library(lubridate)  #Used for time series
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(RColorBrewer)    #Used for color paletting in R
```

Let's check this data and do some cleaning.

```
df <- read_csv("~/Downloads/2006-2012.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   data_id = col_double(),
##   iso = col_double(),
##   event_id_no_cnty = col_double(),
##   year = col_double(),
##   time_precision = col_double(),
##   interaction = col_double(),
##   fatalities = col_double()
```

```
## )
```

```
## See spec(...) for full column specifications.
```

```
names(df)
```

```
##  [1] "X1"               "data_id"          "iso"
##  [4] "event_id_cnty"    "event_id_no_cnty" "event_date"
##  [7] "year"             "time_precision"   "event_type"
## [10] "sub_event_type"   "actor1"           "assoc_actor_1"
## [13] "actor2"           "assoc_actor_2"    "interaction"
## [16] "region"           "country"          "admin1"
## [19] "admin2"           "admin3"           "location"
## [22] "source"           "source_scale"     "notes"
## [25] "fatalities"       "iso3"
```

```
class(df)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
str(df)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 57695 obs. of  26 variables:
##  $ X1              : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ data_id         : num  5747423 5271879 5288547 5805036 5757795 ...
##  $ iso             : num  12 404 566 706 710 729 716 586 586 586 ...
##  $ event_id_cnty   : chr  "ALG3168" "KEN3542" "NIG4288" "SOM10487" ...
##  $ event_id_no_cnty: num  3168 3542 4288 10487 3380 ...
##  $ event_date      : chr  "31-Dec-12" "31-Dec-12" "31-Dec-12" "31-Dec-12" ...
##  $ year            : num  2012 2012 2012 2012 2012 ...
##  $ time_precision  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ event_type      : chr  "Protests" "Violence against civilians" "Riots" "Riots" ...
##  $ sub_event_type  : chr  "Peaceful protest" "Attack" "Violent demonstration" "Violent demonstration"
##  $ actor1          : chr  "Protesters (Algeria)" "Mungiki Militia" "Rioters (Nigeria)" "Rioters (Soma
##  $ assoc_actor_1   : chr  NA NA NA NA ...
##  $ actor2          : chr  NA "Civilians (Kenya)" "Military Forces of Nigeria (1999-2015)" "Rioters (S
##  $ assoc_actor_2   : chr  NA NA NA NA ...
##  $ interaction     : num  60 37 15 55 57 27 60 60 60 60 ...
##  $ region          : chr  "Northern Africa" "Eastern Africa" "Western Africa" "Eastern Africa" ...
##  $ country         : chr  "Algeria" "Kenya" "Nigeria" "Somalia" ...
##  $ admin1          : chr  "Ouargla" "Nyeri" "Plateau" "Awdal" ...
##  $ admin2          : chr  "Ouargla" "Nyeri Town" "Jos South" "Borama" ...
##  $ admin3          : chr  NA "Rware" NA NA ...
##  $ location        : chr  "Ouargla" "Nyeri" "Barakin Kuru Baba" "Borama" ...
##  $ source          : chr  "Al Jazeera" "Star (Kenya)" "Daily Leadership (Nigeria)" "Local Source" ...
##  $ source_scale    : chr  "National" "National" "Subnational" "Other" ...
##  $ notes           : chr  "Unemployed continue protests in Algerias southern region of Ouargla" "Mung
##  $ fatalities      : num  0 0 7 1 0 0 0 0 0 0 ...
##  $ iso3            : chr  "DZA" "KEN" "NGA" "SOM" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_double(),
##   ..   data_id = col_double(),
##   ..   iso = col_double(),
##   ..   event_id_cnty = col_character(),
##   ..   event_id_no_cnty = col_double(),
##   ..   event_date = col_character(),
```

```
##   ..   year = col_double(),
##   ..   time_precision = col_double(),
##   ..   event_type = col_character(),
##   ..   sub_event_type = col_character(),
##   ..   actor1 = col_character(),
##   ..   assoc_actor_1 = col_character(),
##   ..   actor2 = col_character(),
##   ..   assoc_actor_2 = col_character(),
##   ..   interaction = col_double(),
##   ..   region = col_character(),
##   ..   country = col_character(),
##   ..   admin1 = col_character(),
##   ..   admin2 = col_character(),
##   ..   admin3 = col_character(),
##   ..   location = col_character(),
##   ..   source = col_character(),
##   ..   source_scale = col_character(),
##   ..   notes = col_character(),
##   ..   fatalities = col_double(),
##   ..   iso3 = col_character()
##   .. )
```

#I want to remove some columns which only for data entry

```
df1 <- df[-c(2, 3, 4, 5, 8,12,14, 26)]
names(df1)
```

```
##  [1] "X1"             "event_date"     "year"           "event_type"
##  [5] "sub_event_type" "actor1"         "actor2"         "interaction"
##  [9] "region"         "country"        "admin1"         "admin2"
## [13] "admin3"         "location"       "source"         "source_scale"
## [17] "notes"          "fatalities"
```

```
#Looking at columns such as actor1, admin1, notes and so on, we see that they have some missing data.
apply(df1, 2, function(x) any(is.na(x)))
```

```
##             X1     event_date           year     event_type sub_event_type
##          FALSE          FALSE          FALSE          FALSE          FALSE
##         actor1         actor2    interaction         region        country
##          FALSE           TRUE          FALSE          FALSE          FALSE
##         admin1         admin2         admin3       location         source
##          FALSE           TRUE           TRUE          FALSE          FALSE
##   source_scale          notes     fatalities
##          FALSE           TRUE          FALSE
```

```
df1$actor2[is.na(df1$actor2)] <- "NONE"
head(df1$actor2)
```

```
## [1] "NONE"
## [2] "Civilians (Kenya)"
## [3] "Military Forces of Nigeria (1999-2015)"
## [4] "Rioters (Somalia)"
## [5] "Civilians (South Africa)"
## [6] "Civilians (Sudan)"
```

```
#The event date column is in DD/MM/YYYY format.We assign this to the data set using dmy() function.
df1$event_date <- dmy(df1$event_date)
```

```r
#for further analysis and aggregating data on monthly basis wew mutate the dataset to add a column for
df1<-df1 %>%
  mutate(MONTH=month(df1$event_date))

head(df1)
```

```
## # A tibble: 6 x 19
##      X1 event_date  year event_type sub_event_type actor1 actor2
##   <dbl> <date>     <dbl> <chr>      <chr>          <chr>  <chr>
## 1     1 2012-12-31  2012 Protests   Peaceful prot~ Prote~ NONE
## 2     2 2012-12-31  2012 Violence ~ Attack         Mungi~ Civil~
## 3     3 2012-12-31  2012 Riots      Violent demon~ Riote~ Milit~
## 4     4 2012-12-31  2012 Riots      Violent demon~ Riote~ Riote~
## 5     5 2012-12-31  2012 Riots      Mob violence   Riote~ Civil~
## 6     6 2012-12-31  2012 Violence ~ Attack         JEM: ~ Civil~
## # ... with 12 more variables: interaction <dbl>, region <chr>,
## #   country <chr>, admin1 <chr>, admin2 <chr>, admin3 <chr>,
## #   location <chr>, source <chr>, source_scale <chr>, notes <chr>,
## #   fatalities <dbl>, MONTH <dbl>
```

```r
df1%>%
  group_by(country)%>%
  summarise(sum(fatalities))%>%
  arrange(desc(`sum(fatalities)`))%>%
  head(10)
```

```
## # A tibble: 10 x 2
##    country                       `sum(fatalities)`
##    <chr>                                     <dbl>
##  1 Sudan                                     16231
##  2 Democratic Republic of Congo              13221
##  3 Somalia                                   12118
##  4 Pakistan                                   9669
##  5 Nigeria                                    8736
##  6 Ethiopia                                   7225
##  7 South Sudan                                6740
##  8 Libya                                      3975
##  9 Kenya                                      3390
## 10 Chad                                       3389
```

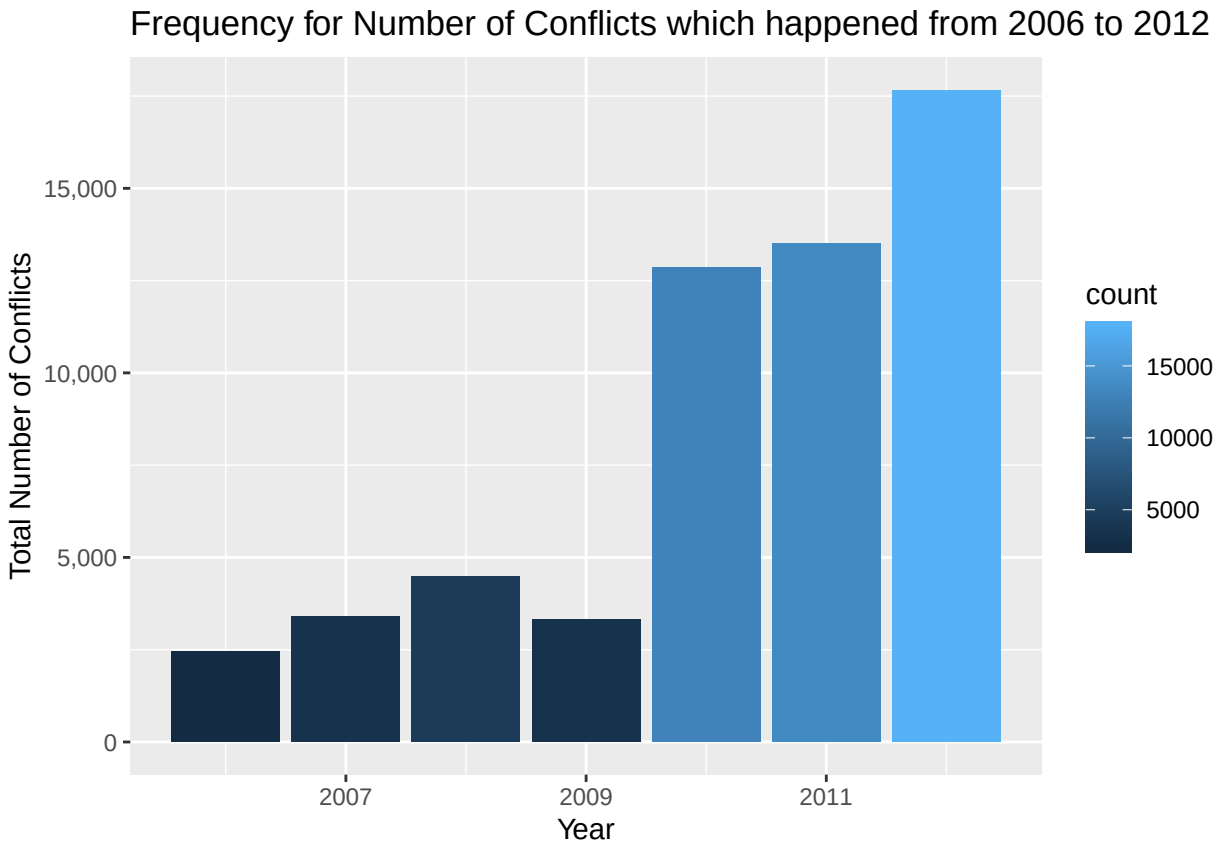
#We see that Sudan has the largest number of fatalities in all other the countries from year 2006 t0 2012.

```r
summary(df1)
```

```
##        X1          event_date              year       event_type
##  Min.   :    1   Min.   :2006-01-01   Min.   :2006   Length:57695
##  1st Qu.:14424   1st Qu.:2010-01-22   1st Qu.:2010   Class :character
##  Median :28848   Median :2011-02-23   Median :2011   Mode  :character
##  Mean   :28848   Mean   :2010-10-06   Mean   :2010
##  3rd Qu.:43272   3rd Qu.:2012-03-03   3rd Qu.:2012
##  Max.   :57695   Max.   :2012-12-31   Max.   :2012
##  sub_event_type        actor1             actor2           interaction
##  Length:57695       Length:57695       Length:57695       Min.   :11.0
##  Class :character   Class :character   Class :character   1st Qu.:16.0
##  Mode  :character   Mode  :character   Mode  :character   Median :37.0
##                                                           Mean   :38.9
```

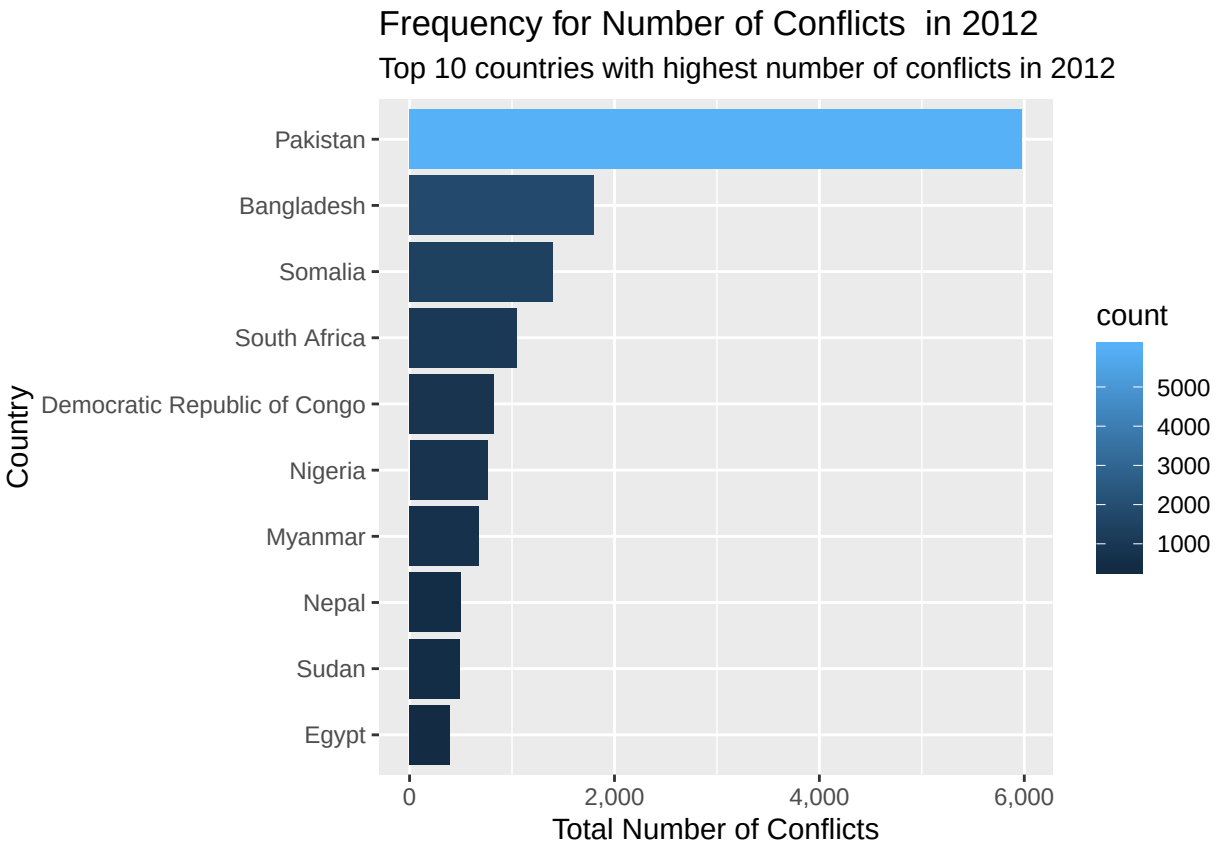
```
##                                                       3rd Qu.:60.0
##                                                       Max.   :88.0
##     region            country             admin1
##  Length:57695       Length:57695       Length:57695
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     admin2             admin3            location
##  Length:57695       Length:57695       Length:57695
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     source          source_scale          notes
##  Length:57695       Length:57695       Length:57695
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##    fatalities           MONTH
##  Min.   :   0.000   Min.   : 1.000
##  1st Qu.:   0.000   1st Qu.: 3.000
##  Median :   0.000   Median : 6.000
##  Mean   :   1.799   Mean   : 6.256
##  3rd Qu.:   1.000   3rd Qu.: 9.000
##  Max.   :1037.000   Max.   :12.000
```

```r
##creating data set for ggplot()
plot <- df1 %>%
  group_by(year)%>%
  summarise(count=n())%>%
  arrange(desc(count))

#using ggplot for visualization
plot %>%
  ggplot(aes(x = year,y=count,fill=count)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(name = "Year") +
  scale_y_continuous(name = "Total Number of Conflicts",labels = scales::comma) +
  ggtitle("Frequency for Number of Conflicts which happened from 2006 to 2012")
```



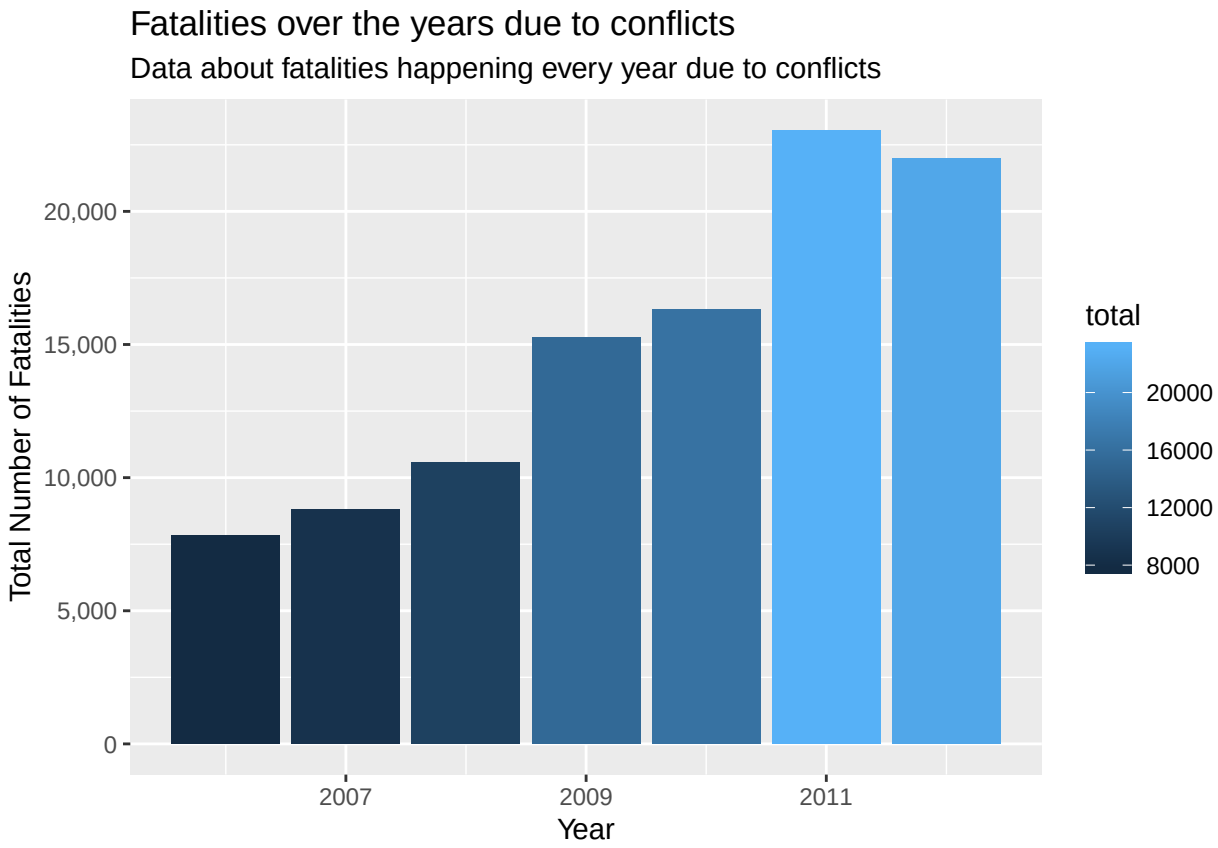
#From this graph, we can see these is a decrese in 2009 and it jumped dramatically since 2010. And the total number of conflicts keeped increasing. We want to look further on the data of 2012 since it is the closest year from now.

```
#creating data set for ggplot()
plot1 <- df1 %>%
  dplyr::filter(year == 2012)%>%
  group_by(country)%>%
  summarise(count=n())%>%
  arrange(desc(count))%>%
  head(10)


  #using ggplot for visualization
plot1 %>%
  arrange(desc(count))%>%
  ggplot(aes(x = reorder(country,count), y=count,fill=count)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(name = "Country") +
  scale_y_continuous(name = "Total Number of Conflicts",labels = scales::comma) +
  ggtitle("Frequency for Number of Conflicts  in 2012",
  subtitle = "Top 10 countries with highest number of conflicts in 2012")+
  coord_flip()
```



As shown, Pakistan is outstandingly observed as the one who has the largest number of conflicts. ## I am interested in investigating the trend of fatalities

```
df1 %>%
  group_by(year)%>%
  summarise(total=sum(fatalities))%>%
  ggplot(aes(x=year,y=total,fill=total))+
  geom_bar(stat = "identity")+
  scale_x_continuous(name = "Year") +
  scale_y_continuous(name = "Total Number of Fatalities",labels = scales::comma) +
  ggtitle("Fatalities over the years due to conflicts",
  subtitle = "Data about fatalities happening every year due to conflicts")
```



This is really a huge data set. And a lot of things can be tested and evaluated from it. I basically used ggplot tool to draw several graphs with the stats I am most interested in. With the given data, I found: 1)Sudan has the largest number of fatalities in all other the countries from year 2006 t0 2012. 2) These is a decrese in 2009 and it jumped dramatically since 2010. And the total number of conflicts keeped increasing. 3)And I dived into the data of 2012, I sadly found that Pakistan is outstandingly observed as the one who has the largest number of conflicts.