EDAV_Final_Project

Akshata Patel, Kiran Saini, Ujjwal Peshin

12/9/2018

1. INTRODUCTION

The issue that we wanted to target was the current migrant and human trafficking incidents around the world. Data on human trafficking and migrant incidents is useful for analysing the trends between countries and developing ways to prevent these crimes of trafficking and exploitation.

1.1 Why we chose this topic.

The primary interest of us choosing this topic to work on was to understand how human trafficking and migration incidents are affecting the world and how big of an issue it is to the world. There are approximately 258 million people not living in their country of birth. Due to the sudden rise in hostility to migration, the illegal routes of human trafficking are becoming more and more prominent.

Some of the questions that we wanted to explore with this data were as follows:

- Which countries have the most fatalities in the migration.
- Which countries face the most illegal trafficking and which countries do people traffick to the most.
- Is there any trend in the victims that are trafficked.
- How have the traffcking trends changed over time. and so on…

1.2 Team Members and contributions:

- Akshata Patel: Akshata worked on the exploratory data analysis of the human trafficking dataset.
- Kiran Saini: Kiran worked on creating the interactive component of the project using D3.
- Ujjwal Peshin: Ujjwal worked on the exploratory data analysis of the migration dataset.

2. DESCRIPTION OF DATA

- The human trafficking dataset used is a global dataset that contains data from counter-trafficking organisations around the world. We have downloaded this data from the CTDC- Counter Trafficking Data Collaborative site[“https://www.ctdatacollaborative.org/ (https://www.ctdatacollaborative.org/)”].

This dataset will be refered to as the human trafficking dataset from here on.

Data Contributors : + International Organization for Migration (IOM) + Polaris + Liberty Asia + Case management services + Counter-trafficking hotline logs

The Global Dataset 3 Sept 2018.csv contains data about the various types of exploitation and means of controls used on the victims. It also includes the country of citizenship of the victim, the country where the exploitation case is registered along with the year of registration. An important column of the dataset is the relationship of the victim with the recuiter. Data collection started in 2017 and the data contains victims registered in the IOM database from years 2002 to 2018. Each type of exploitation, labour performed by the victims and each means of control is an individual column that contains boolean values: 1 for true, 0 for false and -99 for missing values. There are four industries where exploitation based on gender is monitored : Agriculture, construction, manufacturing, domestic.

- The Missing migrants dataset used is a dataset that tracks the death of migrants, or migrants who might have gone missing on route to an international destination. The data can be found at the page of the Missing Migrants Project ["http://missingmigrants.iom.int/downloads (http://missingmigrants.iom.int/downloads)"].

This datset will be refered to as the missing migrants dataset from here on.

Data Contributors: + International Organization for Migration (IOM) + United Nations High Commissioner for Refugees (UNCHR) + Regional Mixed Migration Secretariat + International Red Cross/Red Crescent + NGOs and News Sources

The dataset contains information on migrants who have died or gone missing on a migration route. It inclues all types of incidents, for example, migrants who die in transportation accidents, shipwrecks, violent attacks, or medical complications. It also includes bodies of people identified as migrants who are found at the border of a foreign country. The dataset includes various sources, some of which are media reports, NGO reports, field work, surveys, interviews, reports from the government.

3. ANALYSIS OF DATA QUALITY

3.1 The Human Trafficking dataset

The data includes trafficked people from 2002 to 2018.

```
paste(min(data$yearOfRegistration), "to", max(data$yearOfRegistration))
```

```
## [1] "2002 to 2018"
```

Let us look at the structure of the dataset,

```
str(data)
```

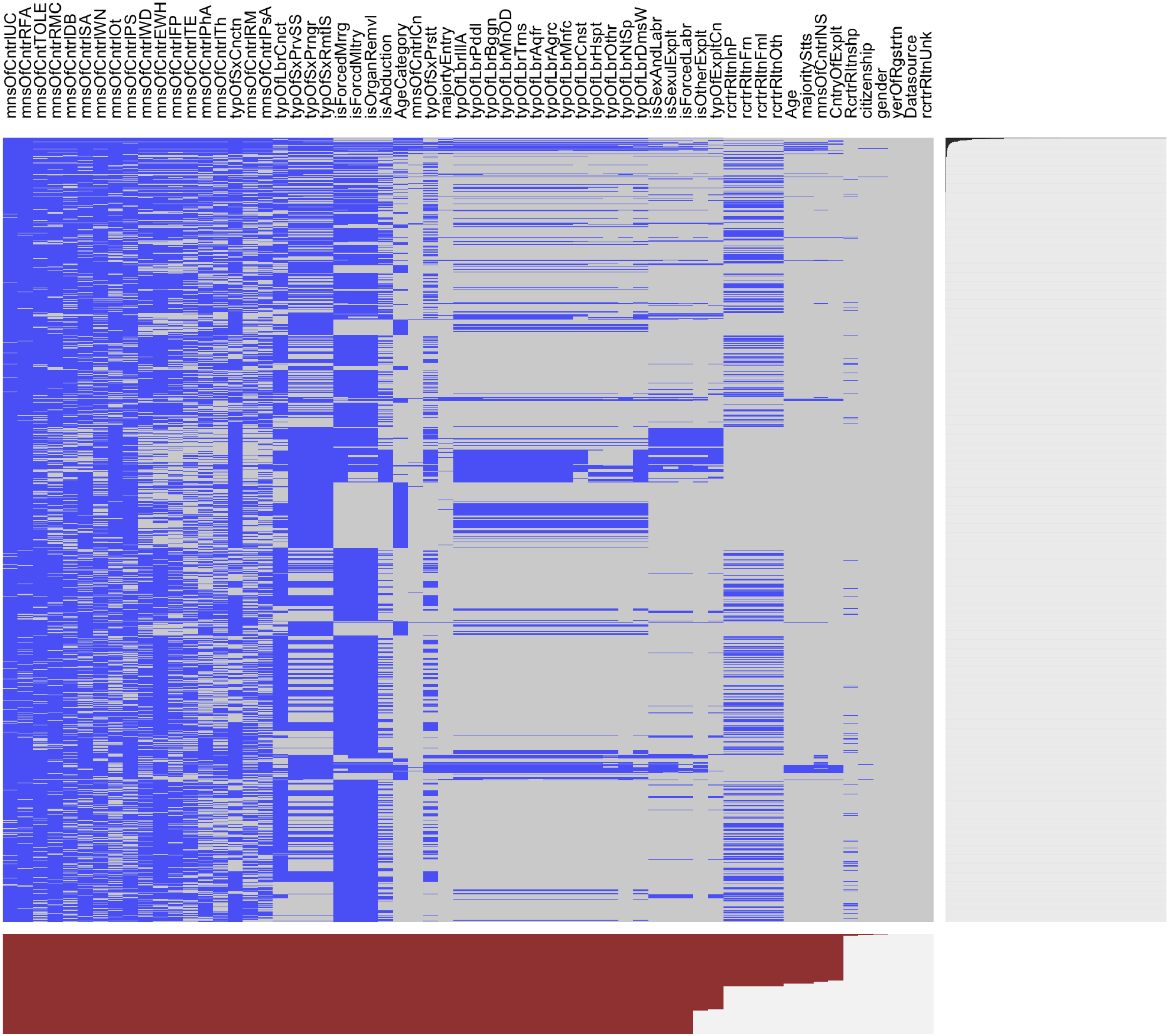
```
## 'data.frame':    55434 obs. of  62 variables:
##  $ yearOfRegistration                    : int  2002 2002 2002 2002 2002 2002 2002 2002 2002 2002 ...
##  $ Datasource                            : Factor w/ 2 levels "Case Management",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ gender                                : Factor w/ 3 levels "Female","Male",..: NA NA NA NA NA NA NA NA NA N
A ...
##  $ Age                                   : Factor w/ 10 levels "0--8","18--20",..: NA NA NA NA NA NA NA NA NA
NA ...
##  $ majorityStatus                        : Factor w/ 3 levels "Adult","Minor",..: NA NA NA NA NA NA NA NA NA N
A ...
##  $ AgeCategory                           : Factor w/ 3 levels "Adult","Minor",..: NA NA NA NA NA NA NA NA NA N
A ...
##  $ majorityEntry                         : Factor w/ 3 levels "Adult","Minor",..: NA NA NA NA NA NA NA NA NA N
A ...
##  $ citizenship                           : Factor w/ 46 levels "AF","AL","BD",..: NA NA NA NA NA NA NA NA NA N
A ...
##  $ meansOfControlDebtBondage             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlTakesEarnings           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlRestrictsFinancialAccess: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlThreats                 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlPsychologicalAbuse      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlPhysicalAbuse           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlSexualAbuse             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlFalsePromises           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlPsychoactiveSubstances  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlRestrictsMovement       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlRestrictsMedicalCare    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlExcessiveWorkingHours   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlUsesChildren            : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlThreatOfLawEnforcement  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlWithholdsNecessities    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlWithholdsDocuments      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlOther                   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlNotSpecified            : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ meansOfControlConcatenated            : Factor w/ 2109 levels "Debt bondage",..: NA NA NA NA NA NA NA NA NA
NA ...
##  $ isForcedLabour                        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ isSexualExploit                       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ isOtherExploit                        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ isSexAndLabour                        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ isForcedMarriage                      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ isForcedMilitary                      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ isOrganRemoval                        : int  NA NA NA NA NA NA NA NA NA NA ...
```

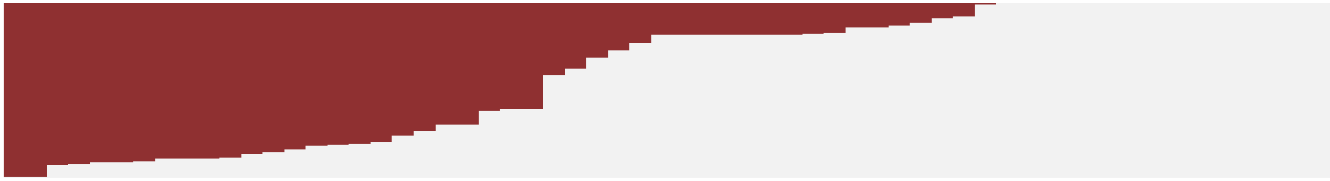
```
## $ typeOfExploitConcatenated : Factor w/ 7 levels "Forced labour",..: NA NA NA NA NA NA NA NA NA N
A ...
## $ typeOfLabourAgriculture : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourAquafarming : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourBegging : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourConstruction : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourDomesticWork : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourHospitality : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourIllicitActivities : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourManufacturing : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourMiningOrDrilling : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourPeddling : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourTransportation : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourOther : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourNotSpecified : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfLabourConcatenated : Factor w/ 17 levels "Agriculture",..: NA NA NA NA NA NA NA NA NA NA
...
## $ typeOfSexProstitution : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfSexPornography : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfSexRemoteInteractiveServices : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfSexPrivateSexualServices : int NA NA NA NA NA NA NA NA NA NA ...
## $ typeOfSexConcatenated : Factor w/ 3 levels "Pornography",..: NA NA NA NA NA NA NA NA NA NA
...
## $ isAbduction : int NA NA NA NA NA NA NA NA NA NA ...
## $ RecruiterRelationship : Factor w/ 17 levels "Family/Relative",..: 17 17 17 17 17 17 17 17 1
7 17 ...
## $ CountryOfExploitation : Factor w/ 57 levels "AE","AF","AL",..: NA NA NA NA NA NA NA NA NA N
A ...
## $ recruiterRelationIntimatePartner : int 0 0 0 0 0 0 0 0 0 0 ...
## $ recruiterRelationFriend : int 0 0 0 0 0 0 0 0 0 0 ...
## $ recruiterRelationFamily : int 0 0 0 0 0 0 0 0 0 0 ...
## $ recruiterRelationOther : int 0 0 0 0 0 0 0 0 0 0 ...
## $ recruiterRelationUnknown : int 1 1 1 1 1 1 1 1 1 1 ...
```

3.1.1 Missing Data

There are a lot of NAs in our data, so let us check the trends of NAs in our data,

```
extracat::visna(data, sort = 'b')
```


mnsOfCntrlUC
mnsOfCntrRFA
mnsOfCntTOLE
mnsOfCntrRMC
mnsOfCntrlDB
mnsOfCntrlSA
mnsOfCntrlWN
mnsOfCntrlOt
mnsOfCntrlPS
mnsOfCntrlWD
mnsOfCntrEWH
mnsOfCntrlFP
mnsOfCntrlTE
mnsOfCntrlPhA
mnsOfCntrlTh
typOfSxCnctn
mnsOfCntrlRM
mnsOfCntrlPsA
typOfLbrCnct
typOfSxPrvSS
typOfSxPrngr
typOfSxRmtlS
isForcedMrrg
isForcdMltry
isOrganRemvl
isAbduction
AgeCategory
mnsOfCntrlCn
typOfSxPrstt
majortyEntry
typOfLbrllA
typOfLbrPddl
typOfLbrBggn
typOfLbrMnOD
typOfLbrTrns
typOfLbrAqfr
typOfLbrAgrc
typOfLbrMnfc
typOfLbrCnst
typOfLbrHspt
typOfLbrOthr
typOfLbrNtSp
typOfLbrDmsW
isSexAndLabr
isSexulExplt
isForcedLabr
isOtherExplt
typOfExpltCn
rcrtrRltnlnP
rcrtrRltnFm
rcrtrRltnFml
rcrtrRltnOth
Age
majoritySttts
mnsOfCntrlNS
CntryOfExplt
RcrtrRltnshp
citizenship
gender
yerOfRgstrtn
Datasource
rcrtrRltnUnk



3.1.2 Unequal Age Intervals

The Age column gives the age of the victim at the time of the exploitation. It contains levels:

```
unique(data$Age)
```

```
##  [1] <NA>    18--20  21--23  24--26  27--29  30--38  9--17   0--8
##  [9] 39--47  48+     Unknown
## 10 Levels: 0--8 18--20 21--23 24--26 27--29 30--38 39--47 48+ ... Unknown
```

3.1.3 The levels of the Age column are unequal intervals. This data makes the plots of Age confusing because even if there are less victims of a particular age, the total number of victims in a range would be very high if the range is large,compared to other smaller ranges. This hinders the correct comprehension of the plots. So we created levels with equal Age intervals.

```
data$Age<-factor(data$Age,levels=c("0--8","9--17","18--20","21--23","24--26","27--29","30--38","39--47","48+"))
data_new<-data %>% filter(data$Age!="NA")
create_age_new<-function(ageBroad)
{
if(ageBroad=="0--8"){
  "0--8"
}else if(ageBroad=="9--17"){
  "9--17"
}else if(ageBroad=="18--20"){
  "18--26"
}else if(ageBroad=="21--23"){
  "18--26"
}else if(ageBroad=="24--26"){
  "18--26"
}else if(ageBroad=="27--29"){
  "27--38"
}else if(ageBroad=="30--38"){
  "27--38"
}else if(ageBroad=="39--47"){
  "39--47"
}else if(ageBroad=="48+"){
  "48+"
}
}

data_new$Age<-sapply(data_new$Age,create_age_new)
data_new$Age<-factor(data_new$Age,levels=c("0--8","9--17","18--26","27--38","39--47","48+"))

unique(data_new$Age)
```

```
## [1] 18--26 27--38 9--17  0--8   39--47 48+
## Levels: 0--8 9--17 18--26 27--38 39--47 48+
```

3.1.4 Country Names and ISO Codes

The dataset contains countries with their ISO codes. The plug-in “Datamaps” used for the interactive component requires alpha-3 codes for countries for plotting it on the world map. We calculated the alpha3 codes along with the country latitudes and longitudes.

3.2 The Missing Migrants dataset

The dataset starts from 2014 and extendes to December 2018.

3.2.1 We can have a look at the structure of the dataset,

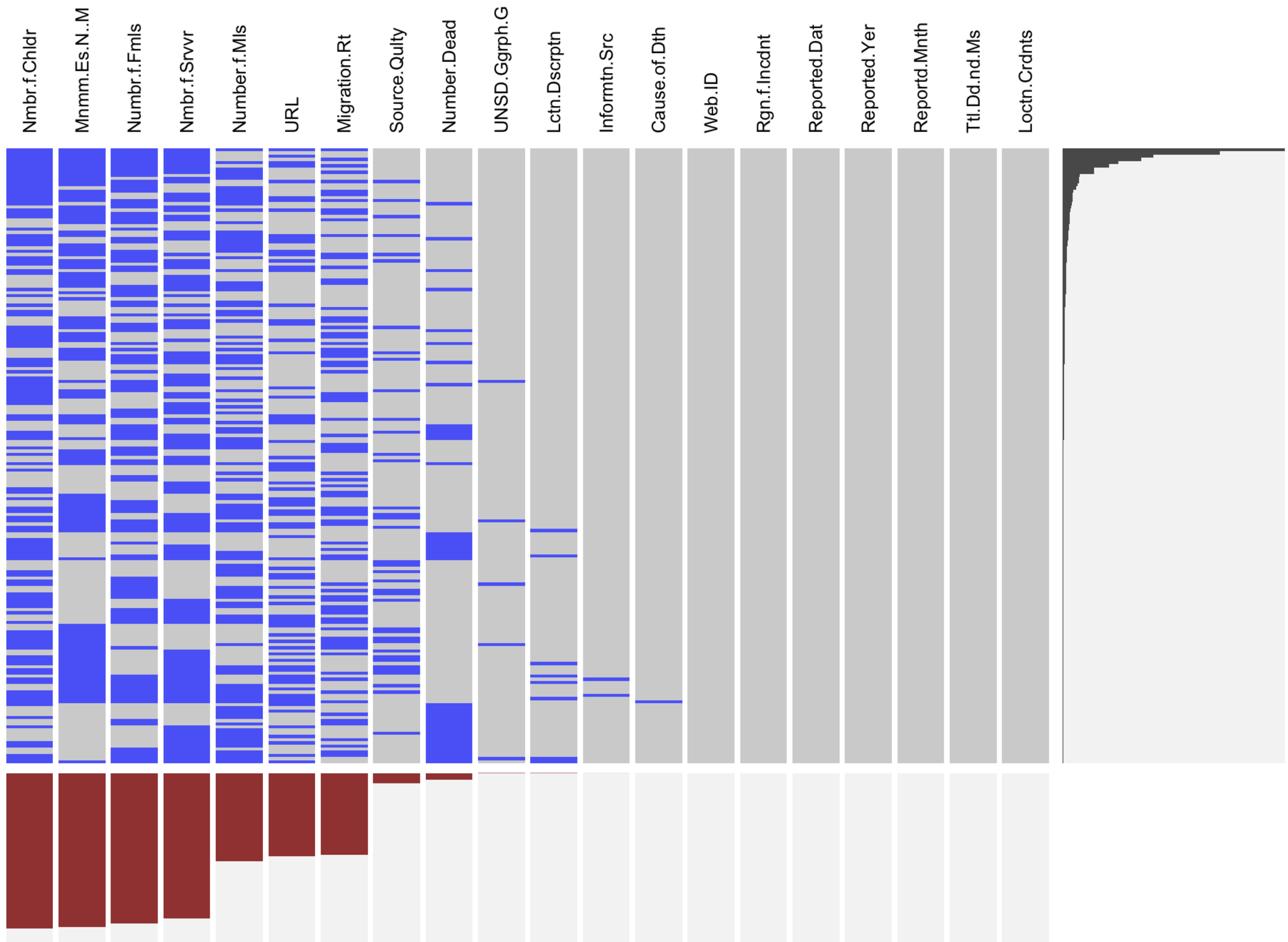
```
str(data)
```

```
## 'data.frame':    4355 obs. of  20 variables:
##  $ Web.ID                             : int  46130 46131 46129 46128 46127 46126 46116 46125 46114 46115 ...
##  $ Region.of.Incident                 : Factor w/ 15 levels "Caribbean","Central America",..: 7 7 15 8 15 15 7
15 15 5 ...
##  $ Reported.Date                      : Factor w/ 1387 levels "April 01, 2015",..: 251 251 247 247 247 243 243
240 1148 1148 ...
##  $ Reported.Year                      : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
##  $ Reported.Month                     : Factor w/ 12 levels "Apr","Aug","Dec",..: 3 3 3 3 3 3 3 3 10 10 ...
##  $ Number.Dead                        : int  12 3 1 3 1 1 2 1 1 1 ...
##  $ Minimum.Estimated.Number.of.Missing: int  3 3 NA NA NA NA NA NA NA NA ...
##  $ Total.Dead.and.Missing             : int  15 6 1 3 1 1 2 1 1 1 ...
##  $ Number.of.Survivors                : int  10 5 NA 1 NA NA 32 NA 1 NA ...
##  $ Number.of.Females                  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Number.of.Males                    : int  NA 2 1 3 NA NA 2 NA 1 1 ...
##  $ Number.of.Children                 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Cause.of.Death                     : Factor w/ 188 levels "Accident (non-vehicle)",..: 105 105 105 89 105 1
05 105 177 84 31 ...
##  $ Location.Description               : Factor w/ 2580 levels " 85 bodies found in Tripoli and 10 near Sabarth
a",..: 1583 1500 1883 2127 1913 1865 1494 1772 1389 679 ...
##  $ Information.Source                 : Factor w/ 1238 levels " El Siglo de Durango",..: 566 507 462 612 1171
1171 50 1171 366 907 ...
##  $ Location.Coordinates               : Factor w/ 2862 levels "-0.023320800000, 14.024647300000",..: 1772 2393
940 2626 944 881 2262 1075 1633 2698 ...
##  $ Migration.Route                    : Factor w/ 15 levels "Calais to United Kingdom",..: 4 15 NA NA NA NA 15
NA NA 14 ...
##  $ URL                                : Factor w/ 1884 levels "http://1.usa.gov/1ktSAmz",..: 1796 1488 1489 17
92 1793 1793 1498 1793 1753 1460 ...
##  $ UNSD.Geographical.Grouping         : Factor w/ 19 levels "Caribbean","Central Africa",..: 16 16 3 18 9 9 16
9 3 15 ...
##  $ Source.Quality                     : int  4 3 3 3 5 5 3 5 3 5 ...
```

3.2.2 The dataset contains a lot of NAs due to the sources of data not having that complete information, with the columns Number.of.Children and Minimum.Estimated.Number.of.Missing having the most missing values. Also, the

most prevalent missing pattern is the pattern where the Number.of.Children, Minimum.Estimated.Number.of.Missing, Number.of.Females, Number.of.Survivors, URL(URL of source), and Migration.Route are missing,

```
extracat::visna(data, sort = 'b')
```



3.2.3 Also, the data contains the coordinates of each incident, but together in a combined form, as (10,-10), and they were split into their own columns,

```
data <- separate(data = data, col = Location.Coordinates, into = c("coord.y", "coord.x"), sep = ",")
```

Some of these coordinates are not valid, ie, latitude greater than 90 or less than -90, or longitude greater than 180 or less than -180,

```
data <- subset(data , coord.y > -90 & coord.y < 90 & coord.x > -180 & coord.y < 180)
```

3.2.4 The dataset, even though it contained the coordinates, it did not contain a mapping from those coordindates to the continents that they belong to, and that was extracted,

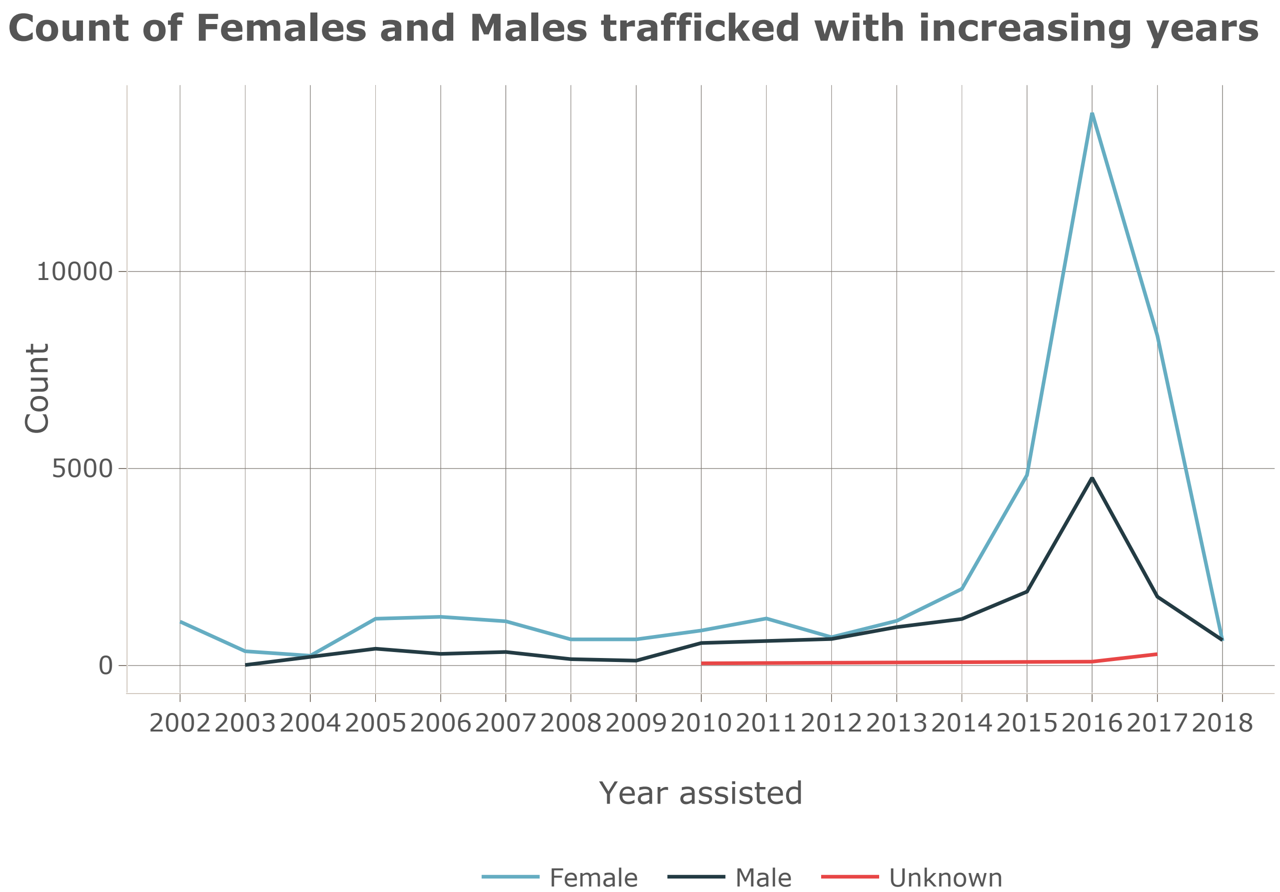
```
coords2continent = function(points)
{
  countriesSP <- getMap(resolution='high')
  pointsSP = SpatialPoints(points, proj4string=CRS(proj4string(countriesSP)))
  indices = over(pointsSP, countriesSP)
  indices$REGION   # returns the continent (7 continent model)
}
```

4. EXPLORATORY DATA ANALYSIS

4.1 The Human Trafficking Dataset

First we check the trends of trafficking with each passing year. We created a line plot of the count of victims with years based on the gender of the victims.

```
data$count <- 1
agg_data <- aggregate(count ~ yearOfRegistration + gender, data = subset(data, !is.na(gender)), FUN = length)
#creating date objects from numeric data
agg_data$yearOfRegistration <- make_date(agg_data$yearOfRegistration)

p <- ggplot() +
      geom_line(data =  agg_data,aes(x= yearOfRegistration, y = count, color = gender)) +
      scale_x_date(date_breaks = "1 year", date_labels = "%Y") +
      theme(legend.position="bottom") +
      xlab("Year assisted") +
      ylab("Count") +
      ggtitle("Count of Females and Males trafficked with increasing years ")

ggplotly(p) %>% layout(legend = list(orientation = "h", y = -0.25, x = 0.30))
```

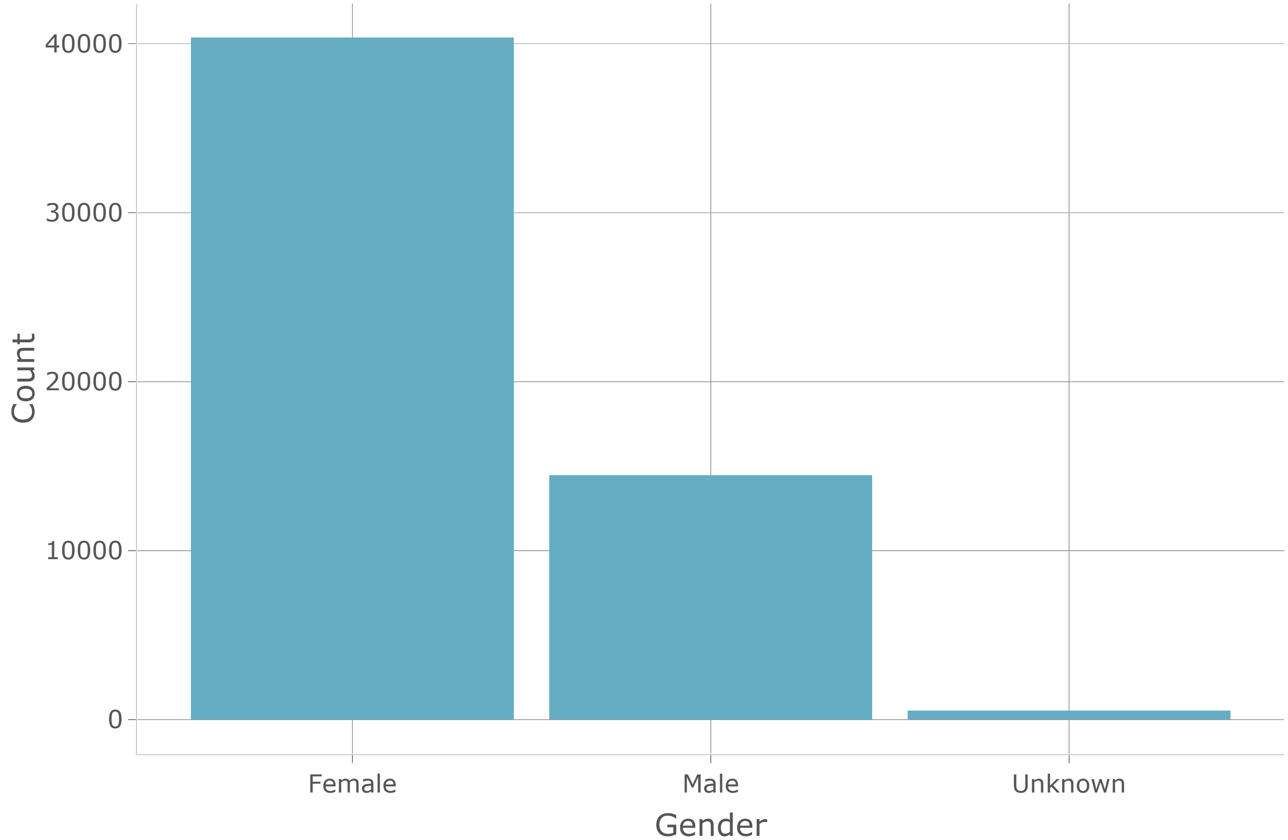


From this graph we see that there was an increase in trafficking cases during the years 2014 to 2017. One can also observe that the number of female victims is greater than male victims for all years.

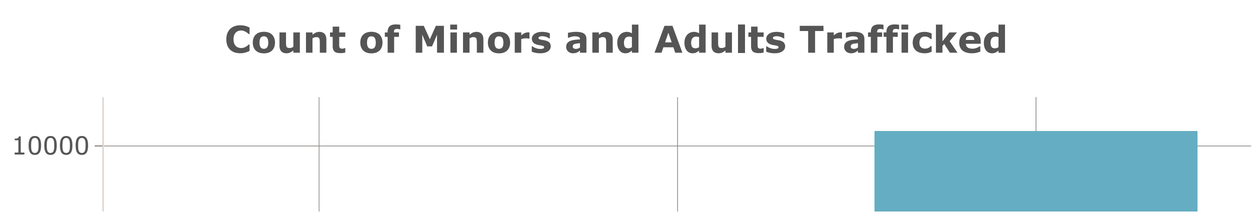
This can be confirmed by the following bar plot that shows the count of victims by gender and Age.

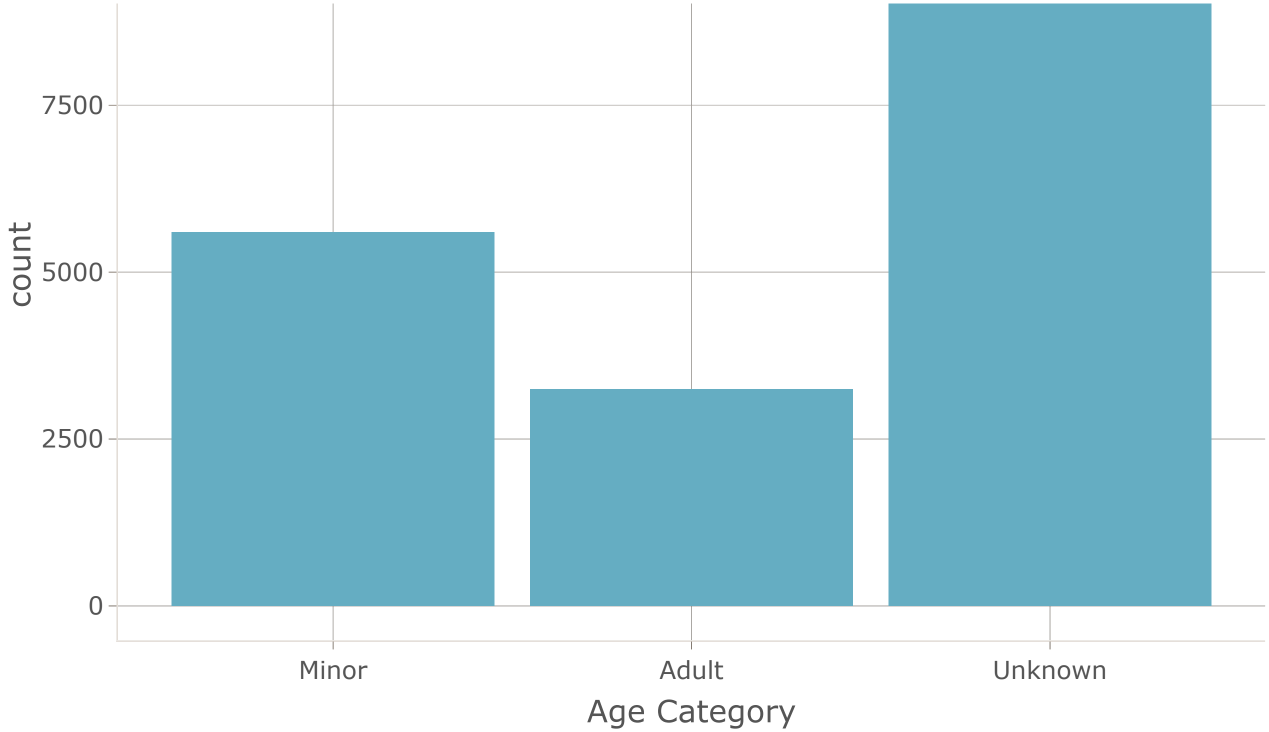
```
p <- ggplot() +
  geom_bar(data = subset(data, !is.na(gender)), aes(gender)) +
  xlab("Gender") +
  ylab("Count") +
  ggtitle("Count of Females and Males Trafficked")

ggplotly(p)
```

Count of Females and Males Trafficked



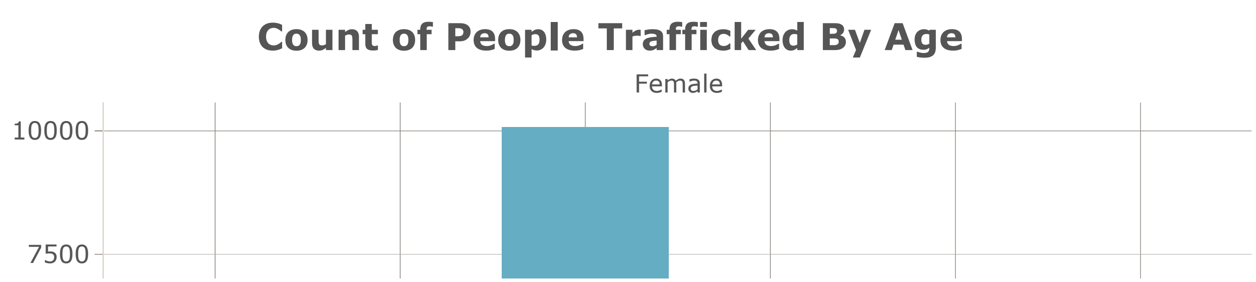

```
f <- factor(data$AgeCategory)
data$AgeCategory <- fct_relevel(f, "Minor")

g <- ggplot() +
  geom_bar(data = subset(data, !is.na(AgeCategory)), aes(AgeCategory)) +
  ggtitle("Count of Minors and Adults Trafficked") +
  xlab("Age Category")

ggplotly(g)
```

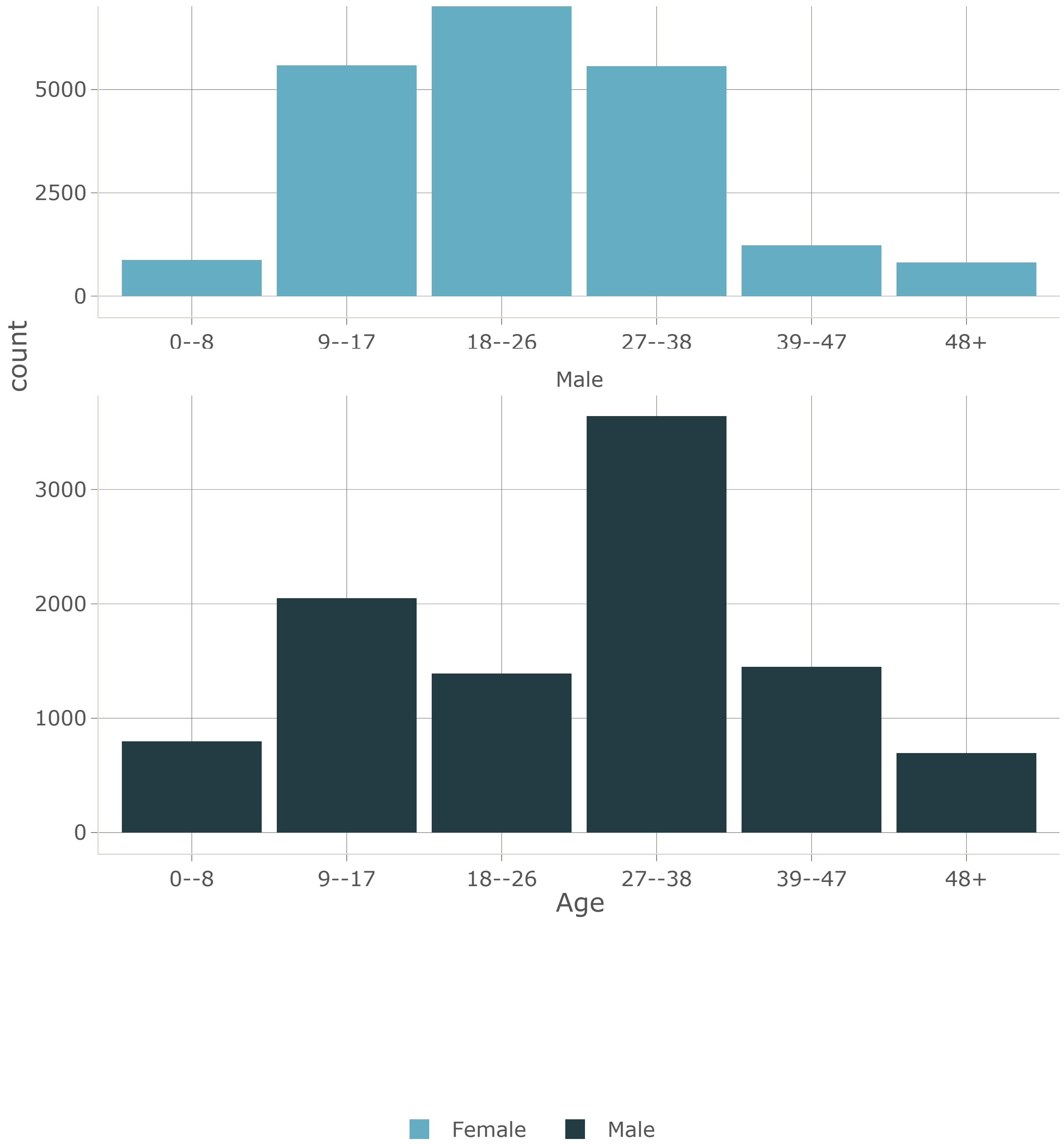




Trafficking victims are mostly females and minors compared to other gender and age groups. A lot of victims' Age Category is unknown. This may be the case because the victims are unaware when the exploitation started, hence they cannot determine the age when the exploitation of the victim began.

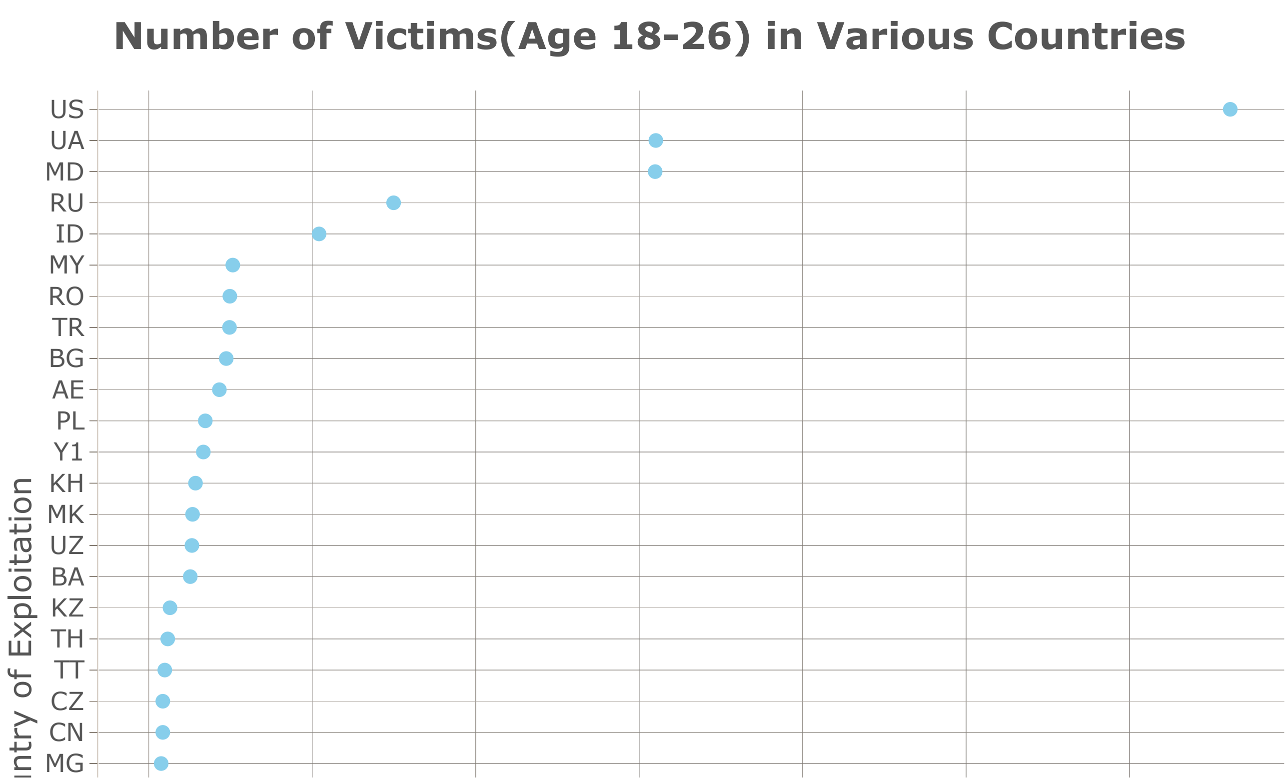
```
p <-  ggplot() +
  geom_bar(data = subset(data_new, !is.na(Age) & gender!="Unknown"), aes(Age, fill =gender), position = "dodge")
 +
  ggtitle("Count of People Trafficked By Age ") +
  facet_wrap(. ~ gender, scales = "free", ncol = 1)

ggplotly(p) %>% layout(legend = list(orientation = "h", y = -0.25, x = 0.30))
```

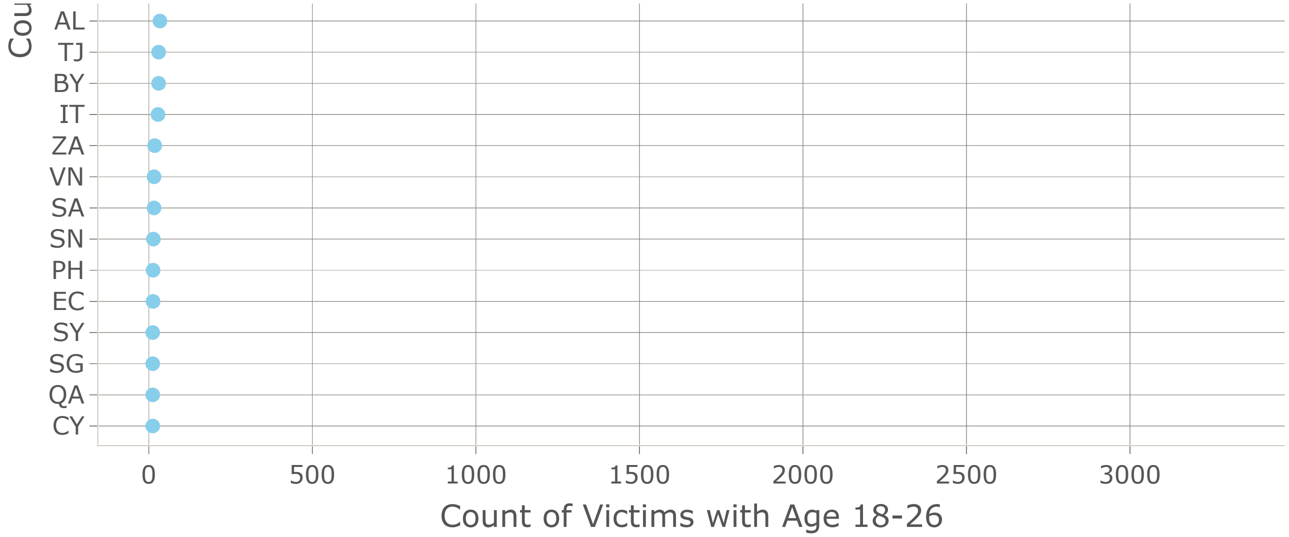



5000
2500
0
0--8
9--17
18--26
27--38
39--47
48+
count
Male
3000
2000
1000
0
0--8
9--17
18--26
27--38
39--47
48+
Age
Female
Male

One can observe that majority of females victims are age group 18-26 and the male victims are of the age 27-38. Feamles are trafficked in their youth whereas the males are trafficked when they have considerable strength to work.

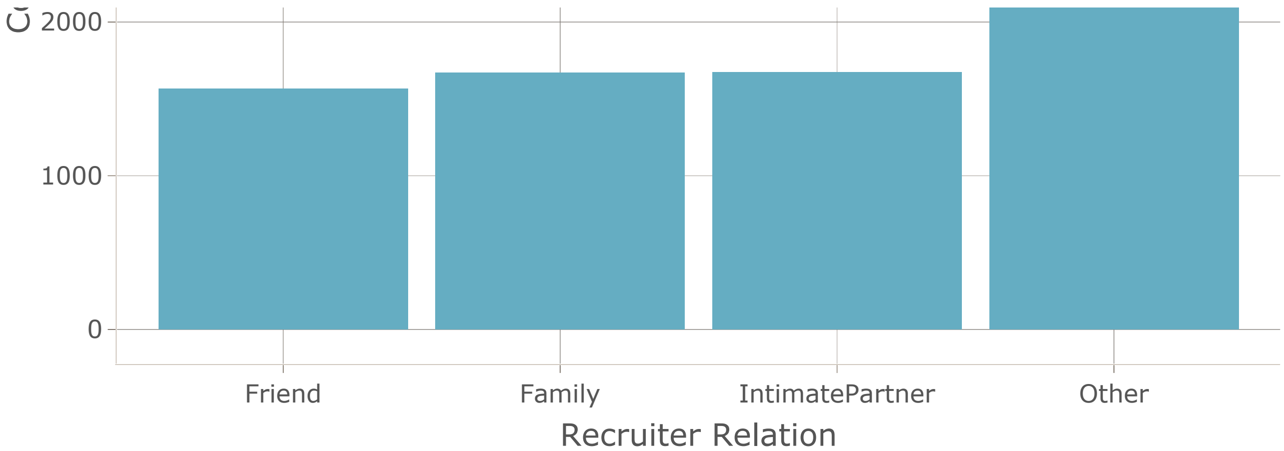
```
data_filtered<-data_new %>% filter(CountryOfExploitation!="-99")
data_filtered<-data_filtered %>% filter(Age=="18--26")
data_count_country <- data_filtered %>% group_by(CountryOfExploitation) %>% summarise(count=n())

Country_Of_Exploitation<-reorder(data_count_country$CountryOfExploitation,data_count_country$count)

p<-ggplot(data_count_country)+
  geom_point(aes(x=Country_Of_Exploitation,y=count),color="skyblue")+
  xlab("Country of Exploitation")+
  ylab("Count of Victims with Age 18-26")+
  scale_y_continuous(breaks=seq(0,4000,500))+
  coord_flip()+
  ggtitle("Number of Victims(Age 18-26) in Various Countries")


ggplotly(p)
```





This cleveland plot shows the count of victims of the age group 18 to 26 with the country of exploitation.

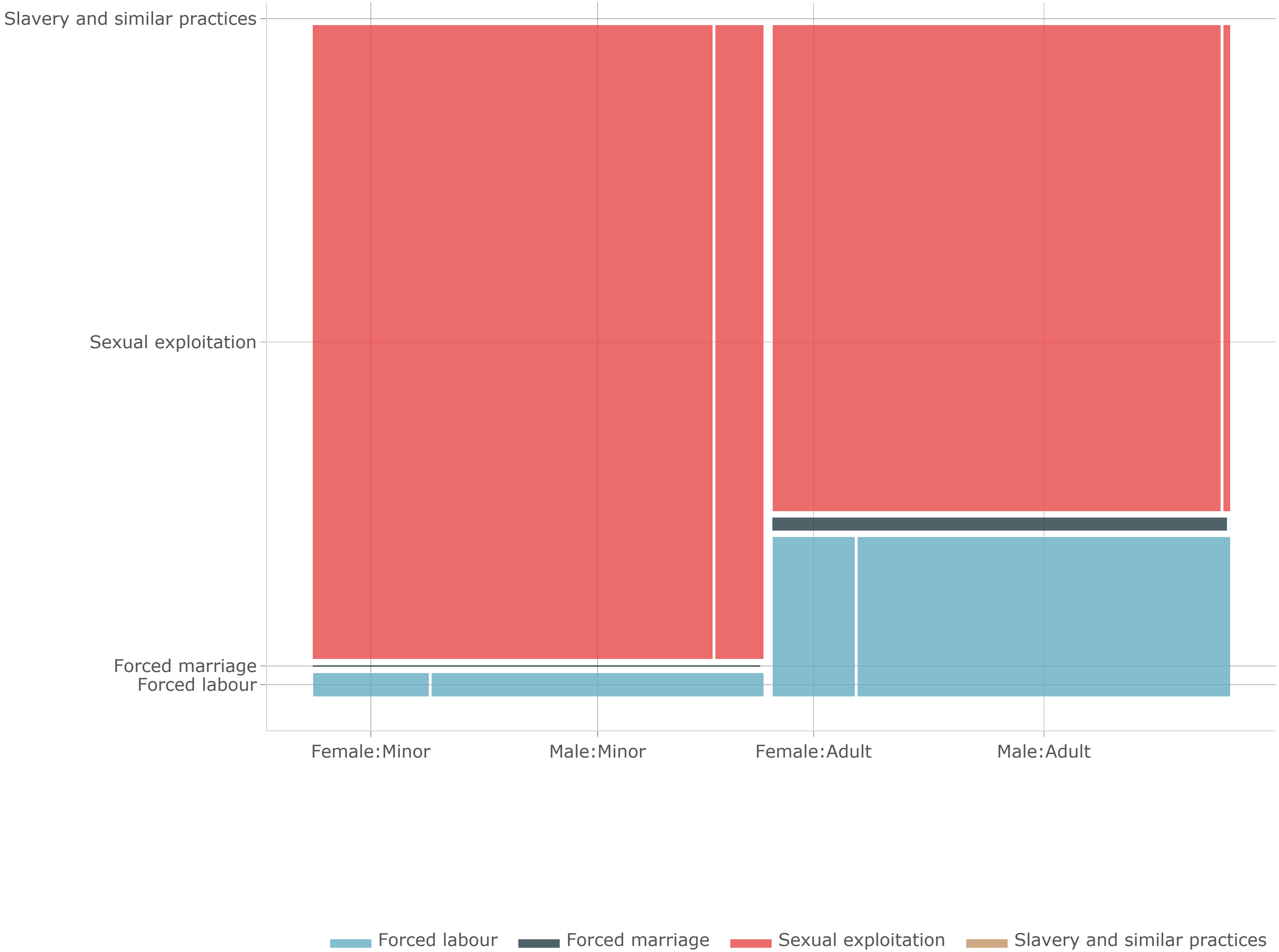
```
small_data <- data[58:61]
small_data <- data.frame(values=colSums(small_data, na.rm=TRUE), names = names(small_data))
small_data$names <- factor(c("Family","Friend","IntimatePartner","Other"))

p <- ggplot() + geom_bar(data = small_data, aes(y = values, x = reorder(names,values)), stat = "identity") +
     ggtitle("Count of Relation of Recruiter with Trafficked Person")+
     xlab("Recruiter Relation") +
     ylab("Count")
ggplotly(p)
```



This shows the relation(Family member, friend, partner, others/unknown person) of the recruiter/trafficker with the victim. We can see that the perpetuator of the crime can be anyone from your family, friends or even your intimate partner.

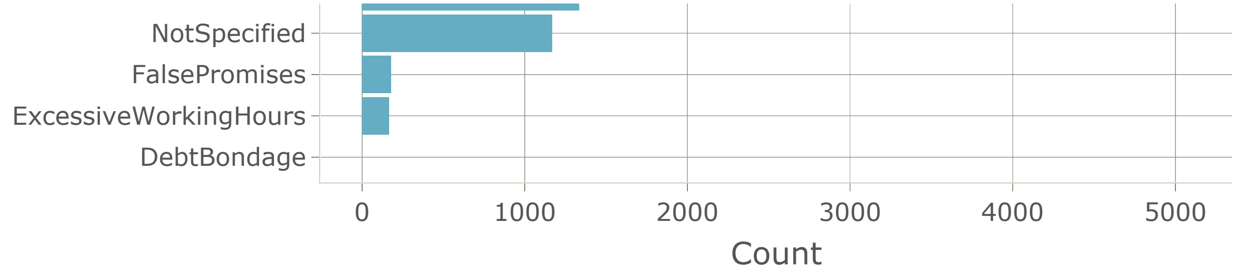
```
data_mosaic <- data %>% filter(typeOfExploitConcatenated!="Forced labour;Sexual exploitation;Combined sexual and
 labour exploitation" & typeOfExploitConcatenated!="Forced labour;Slavery and similar practices" & typeOfExploitC
oncatenated!="Other")

data_mosaic$typeOfExploitConcatenated <- factor(data_mosaic$typeOfExploitConcatenated,levels=c("Forced labour","F
orced marriage","Sexual exploitation","Slavery and similar practices" ))

data_mosaic$gender <- factor(data_mosaic$gender,levels=c("Female","Male") )
data_mosaic$AgeCategory <- factor(data_mosaic$AgeCategory,levels=c("Minor","Adult"))

p<-ggplot(data = data_mosaic) +
   geom_mosaic(aes(x = product(gender,typeOfExploitConcatenated),conds=product(AgeCategory), fill=typeOfExploitCo
ncatenated), na.rm=TRUE) +
  # facet_grid(AgeCategory~.) +
   ggtitle("Type of Exploitation By Gender And Age") +
   xlab("") +
   ylab("") +
   theme(legend.title=element_blank())


ggplotly(p) %>% layout(legend = list(orientation = "h", y = -0.25, x = 0.30))
```

Type of Exploitation By Gender And Age


Slavery and similar practices
Sexual exploitation
Forced marriage
Forced labour
Female:Minor
Male:Minor
Female:Adult
Male:Adult
Forced labour
Forced marriage
Sexual exploitation
Slavery and similar practices

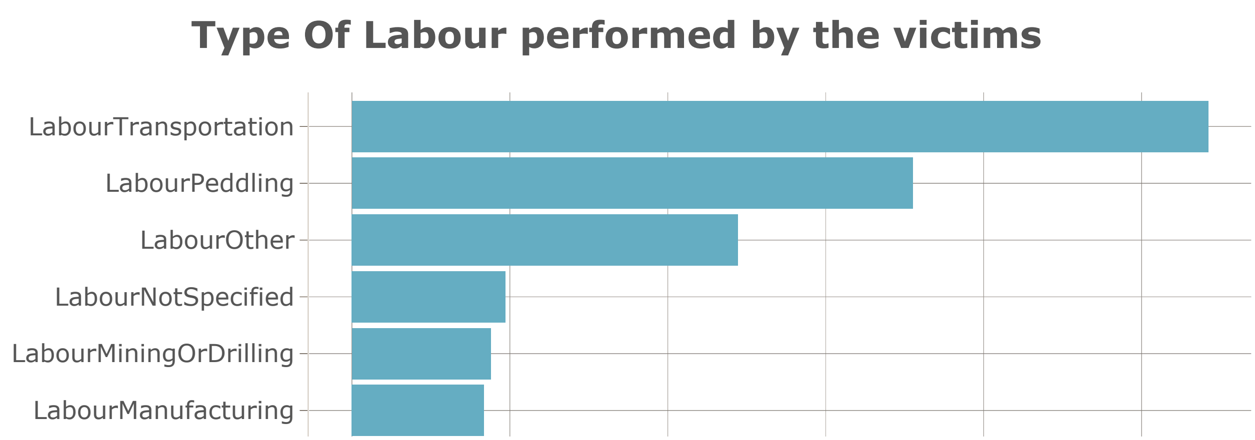
The mosaic plot shows the proportion of the types of exploitation by gender and age category i.e. minor and adult. It can be seen that the proportion of victims sexually exploited is the maximum among both females minors and adults. The proportion of forced labour is greater in male adults than female adults. The proportion of sexual exploitation is greater in female adults than male adults.

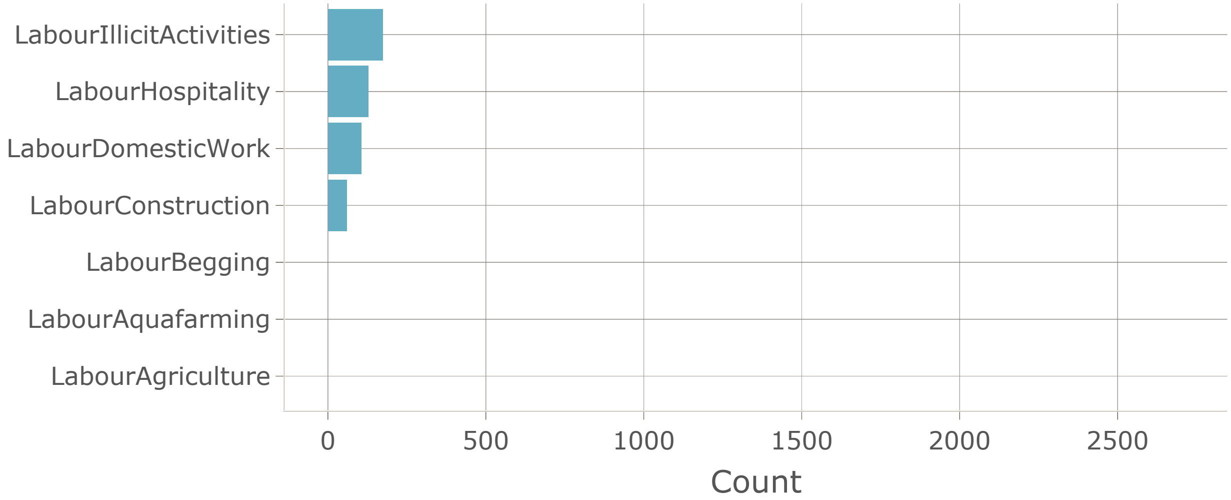
```
small_data <- data[9:26]
small_data <- data.frame(values=colSums(filter(small_data,small_data$meansOfControlNotSpecified!=1), na.rm=TRUE),
names = names(filter(small_data,small_data$meansOfControlNotSpecified!=1)))

small_data <- small_data %>% arrange(values)

factors <- levels(small_data$names)
factors <- sapply(factors,function(one_factor){
  substring(one_factor,15)
})

small_data$names <- factors
ggthemr('fresh')
p <- ggplot() + geom_bar(data = small_data, aes(y = values, x = names), stat = "identity") + coord_flip() +
    ggtitle("Means of control used on Victims") +
    ylab("Count") + xlab("")

ggplotly(p)
```



The bar plot shows the count of victims with the different means of control used by the trafficker on them.

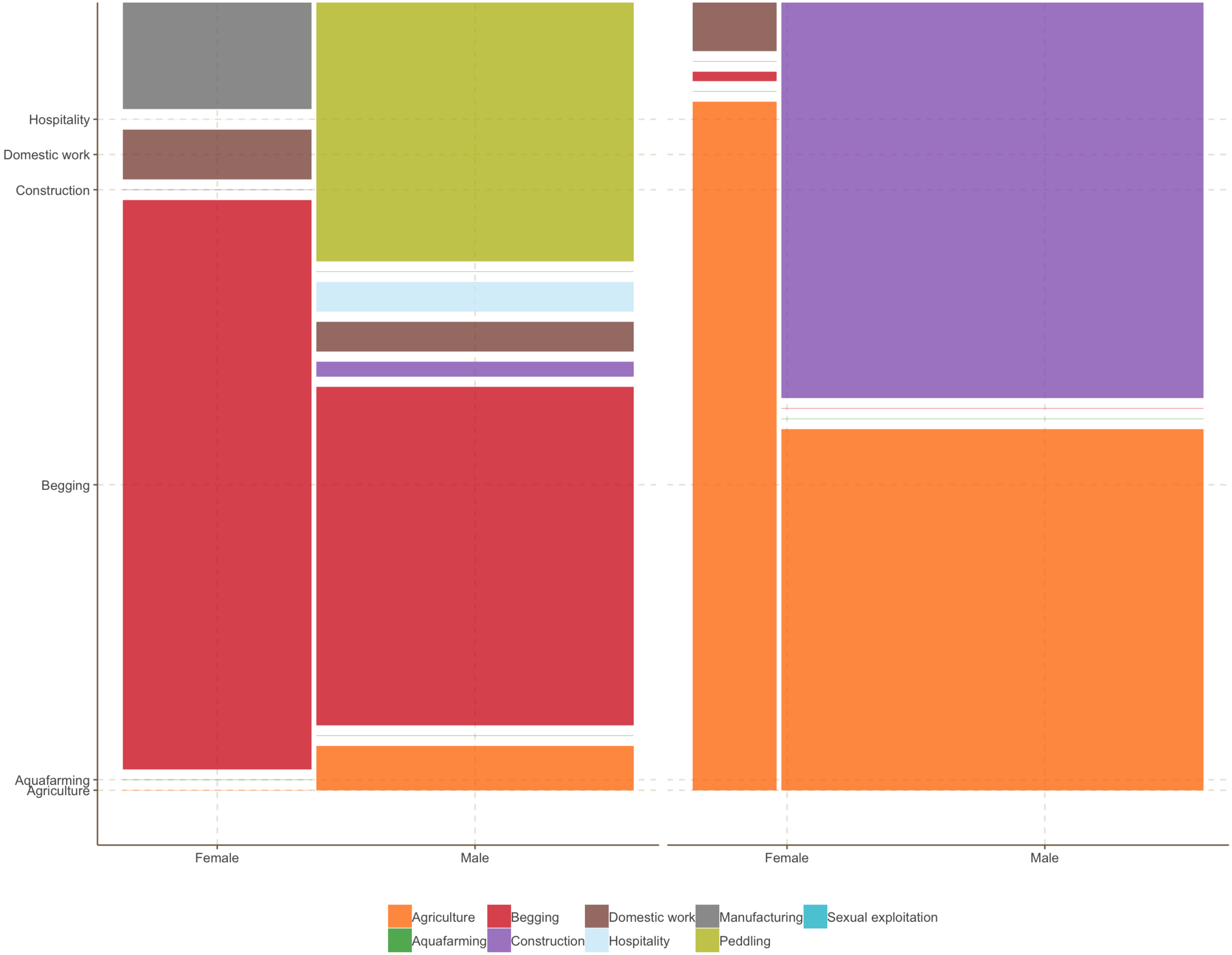
```
small_data <- data[36:48]
small_data <- data.frame(values=colSums(small_data, na.rm=TRUE), names = names(small_data))

small_data <- small_data %>% arrange(values)

factors <- levels(small_data$names)
factors <- sapply(factors,function(one_factor){
  substring(one_factor,7)
})

small_data$names <- factors
ggthemr_reset()
ggthemr('fresh')
p <- ggplot() + geom_bar(data = small_data, aes(y = values, x = names), stat = "identity") + coord_flip() +
    ggtitle("Type Of Labour performed by the victims")+
    ylab("Count") + xlab("") +
    scale_y_continuous(breaks=seq(from = 0, to = 3000, by = 500))
ggplotly(p)
```





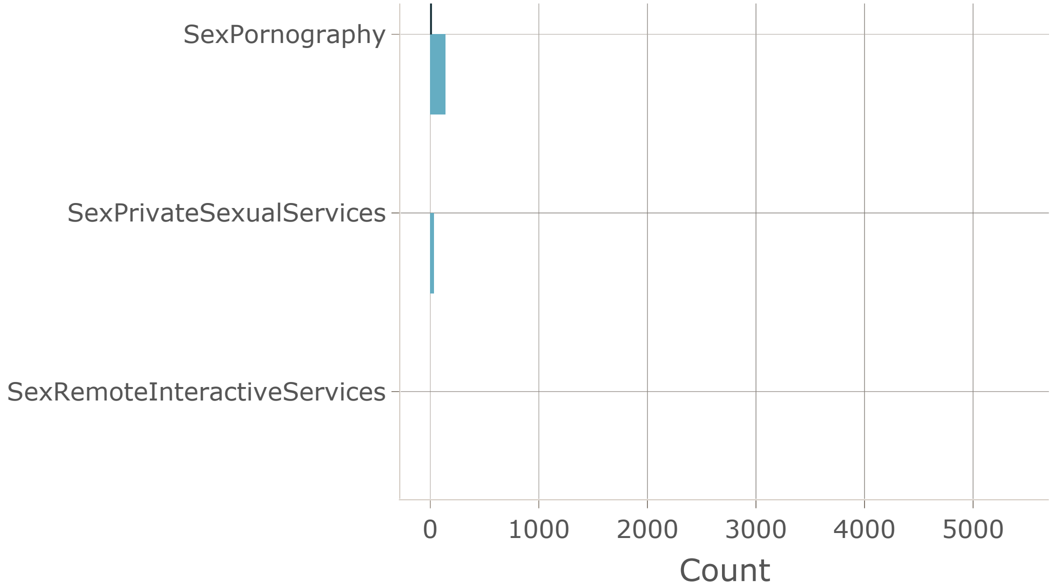
This bar plot shows the count of the victims and the type of labour they are forced to perform. Victims are majorly exploited to work in the labout transportation sector. Types of labour classified by gender and age is shown in the mosaic plot below:

```
data_mosaic <- data %>% filter(typeOfLabourConcatenated!="Agriculture;Not specified" & typeOfLabourConcatenated!=
"Construction;Not specified" & typeOfLabourConcatenated!="Domestic Work;Not specified" & typeOfLabourConcatenate
d!="Domestic work;Other" & typeOfLabourConcatenated!="Other;Not specified" & typeOfLabourConcatenated!="Other" &
 typeOfLabourConcatenated!="Not specified" & AgeCategory!="Unknown")

data_mosaic$typeOfLabourConcatenated <- factor(data_mosaic$typeOfLabourConcatenated,levels=c("Agriculture","Aquaf
arming","Begging","Construction","Domestic work","Hospitality","Manufacturing","Peddling","Sexual exploitation"))

data_mosaic$gender <- factor(data_mosaic$gender,levels=c("Female","Male") )
data_mosaic$AgeCategory <- factor(data_mosaic$AgeCategory,levels=c("Minor","Adult"))

ggthemr('fresh',text_size = 12)
swatch_colours <- c('#1F77B4', '#FF7F0E', '#2CA02C', '#D62728', '#9467BD', '#8C564B', '#CFECF9', '#7F7F7F', '#BCB
D22', '#17BECF')#,"#555555","#EEEEEE")
set_swatch(swatch_colours)

ggplot(data = data_mosaic) +
   geom_mosaic(aes(x = product(typeOfLabourConcatenated , gender), fill=typeOfLabourConcatenated), na.rm=TRUE) +
   facet_grid(.~AgeCategory) +
   ggtitle("Type of Labour Performed By Victim Gender And Age") +
   xlab("") +
   ylab("") +
   theme(legend.title=element_blank()) +
   theme(legend.position="bottom")
```



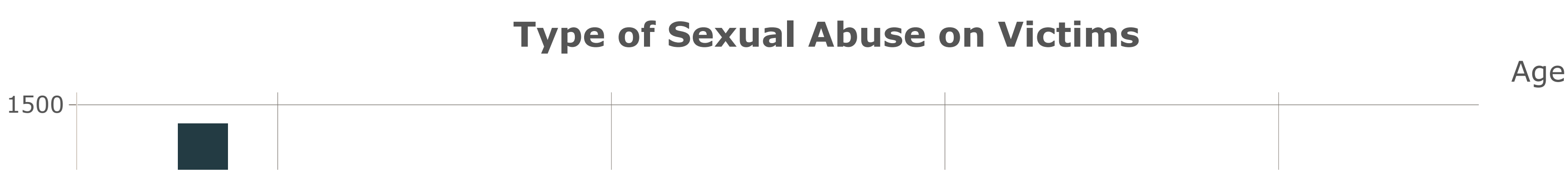
From the graph, one can observe the different types of labour prominent among different age categories of males and females. Minor females are engaged in begging and males are engaged in peddling. Adult females are engaged in agriculture and males (as expected) are employed in construction.

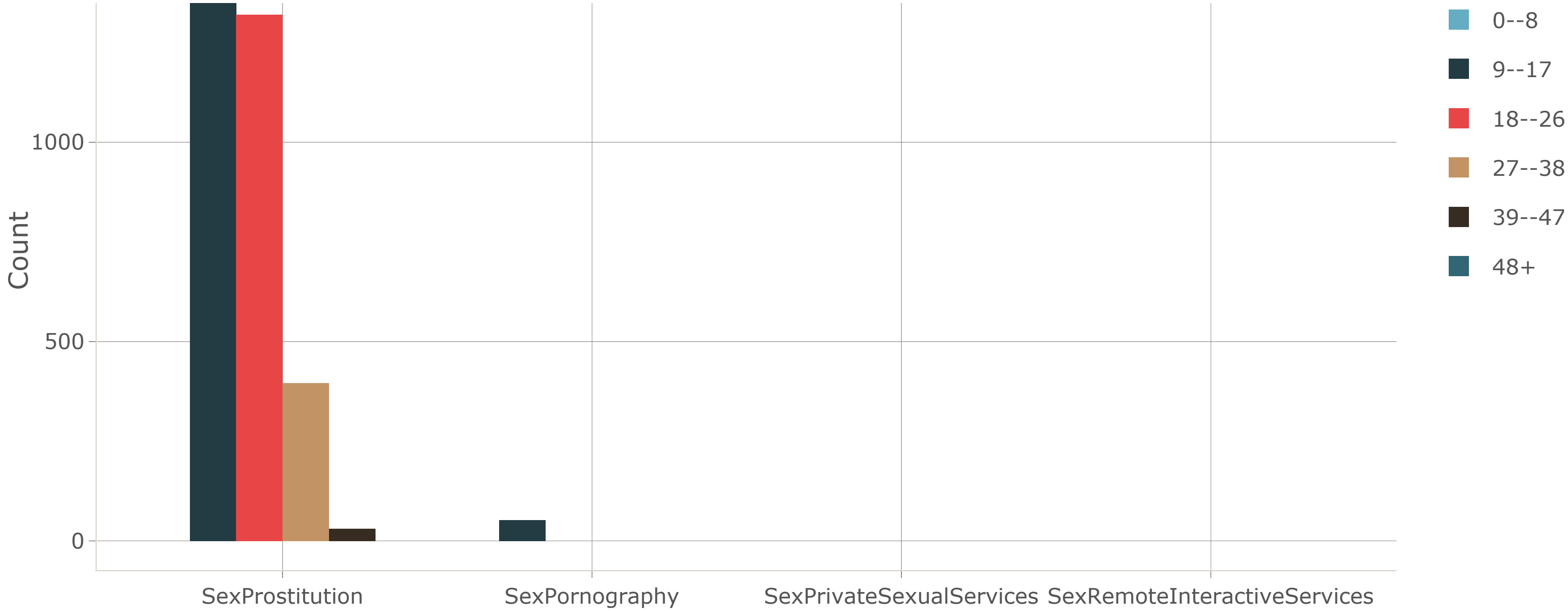
```
small_data <- select(data,c(50:53,3,4))

small_data_female <- small_data %>% filter(gender=="Female") %>% select(1:4)
small_data_female <- data.frame(values=colSums(small_data_female, na.rm=TRUE), names = names(small_data_female),g
ender="Female")


small_data_male <- small_data %>% filter(gender=="Male") %>% select(1:4)
small_data_male <- data.frame(values=colSums(small_data_male, na.rm=TRUE), names = names(small_data_male),gender=
"Male")


#small_data <- data.frame(values=colSums(small_data, na.rm=TRUE), names = names(small_data))
small_data <- rbind(small_data_female,small_data_male)

small_data <- small_data %>% arrange(-values)

factors <- small_data$names
factors <- sapply(factors,function(one_factor){
  substring(one_factor,7)
})

small_data$names <- factors
ggthemr('fresh')
p <- ggplot() + geom_bar(data = small_data, aes(y = values, x = reorder(names,values),fill=gender), stat = "ident
ity",position="dodge") + coord_flip() +
    ggtitle("Type of Sexual Abuse on Victims") +
    ylab("Count") + xlab("")+
    scale_y_continuous(breaks=seq(0,5000,1000))

ggplotly(p)
```



Type of sexual abuse is more prominent for female victims and it is insignificant for male victims. We further saw the distribution of sexual abuse victims' age groups for minor females in the following bar chart.

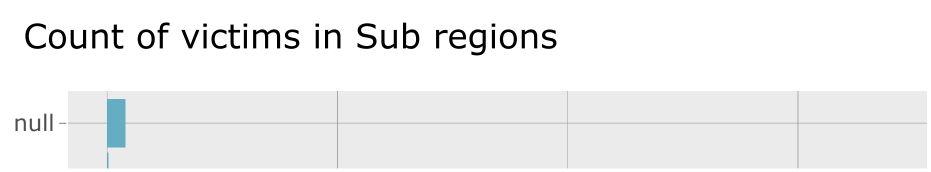
```
small_data_female <- data_new %>% filter(gender=="Female") %>% select(c(50:53,4))

small_data_female <- small_data_female %>% group_by(Age) %>% select(1:5) %>% summarise(SexProstitution= sum(typeOfSexProstitution==1,na.rm=TRUE),SexPornography= sum(typeOfSexPornography==1,na.rm=TRUE),SexPrivateSexualServices= sum(typeOfSexPrivateSexualServices==1,na.rm=TRUE),SexRemoteInteractiveServices= sum(typeOfSexRemoteInteractiveServices==1,na.rm=TRUE))

small_data_female <- gather(small_data_female,key="TypeOfSexualExploitation",value="Value",-Age)
ggthemr('fresh')
p <- ggplot() + geom_bar(data = small_data_female, aes(y=Value,x =reorder(TypeOfSexualExploitation,-Value ),fill=Age), stat = "identity",position="dodge") +
    ggtitle("Type of Sexual Abuse on Victims") +
    ylab("Count") + xlab("")

ggplotly(p)
```

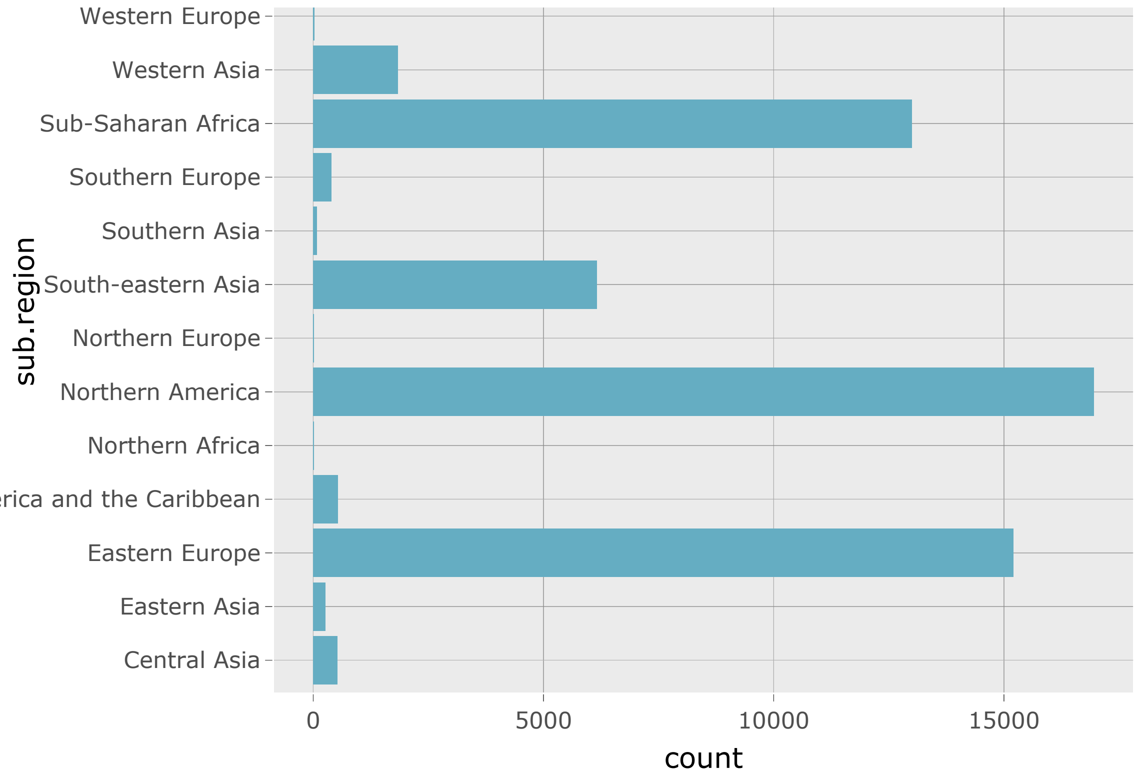




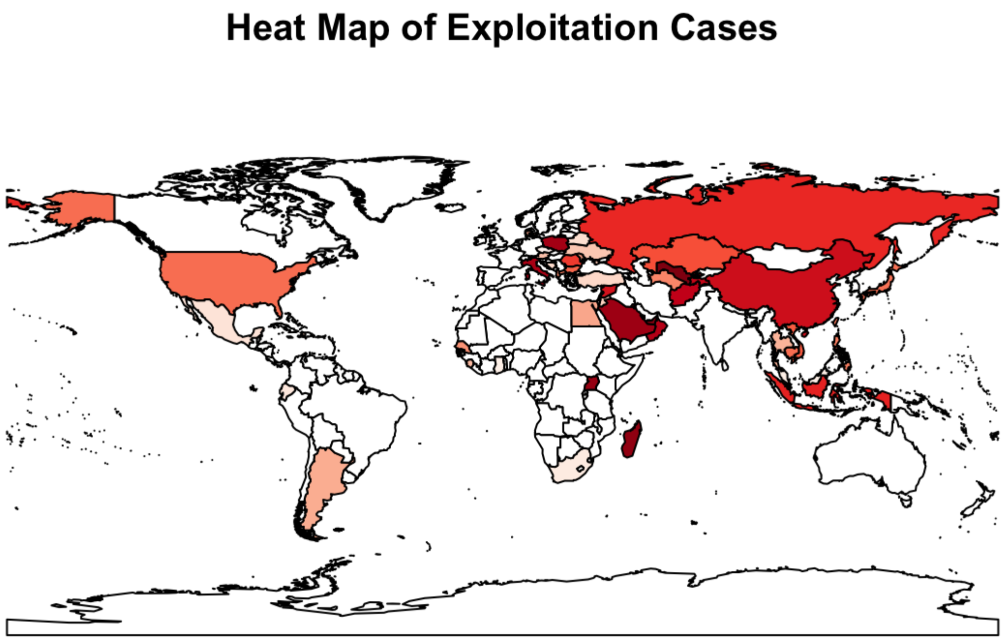
Sexual prostitution is the most common form of sexual abuse in adult females. The majority of these victims are of the age group 9 to 26.

```
countries <- read.csv('./data/all.csv')

data <- merge(x = data,
              y = countries[,c('alpha.2', 'region', 'sub.region')],
              by.x = "CountryOfExploitation",
              by.y = "alpha.2",
              all.x = TRUE)
ggthemr_reset()
p <- ggplot() + geom_bar(data = data, aes(sub.region)) + coord_flip()+
    ggtitle("Count of victims in Sub regions")

ggplotly(p)
```



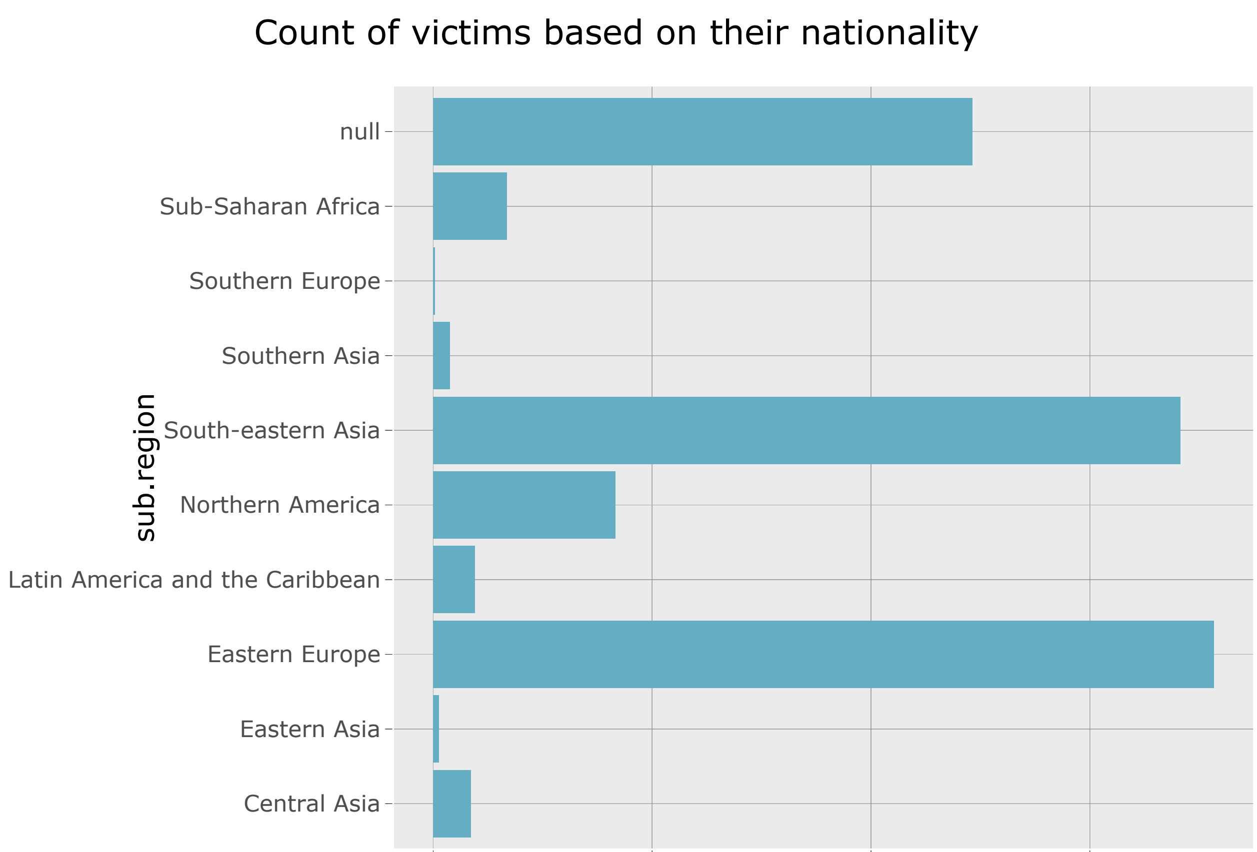



```
data(wrld_simpl)

data_countries <- read.csv("./data/count_country_to_country.csv")

data_countries_sum <- data_countries %>% group_by(CountryOfExploitation) %>% summarise(sum_values = sum(value))

data_countries_sum <- data_countries_sum %>% filter(data_countries_sum$CountryOfExploitation!="NA")

pal <- colorRampPalette(brewer.pal(9, 'Reds'))(length(data_countries_sum$sum_values))
pal <- pal[with(data_countries_sum, findInterval(sum_values, sort(unique(sum_values))))]

col <- rep(grey(1.0), length(wrld_simpl@data$ISO2))

arr<-match(data_countries_sum$CountryOfExploitation, wrld_simpl@data$ISO2)
arr <- arr[!is.na(arr)]
col[arr] <- pal[0:(length(arr))]

plot(wrld_simpl, col = col,main="Heat Map of Exploitation Cases")
```



The world map shows the heat map of the count of cases in the country of exploitation. The darker regions have a higher count of victims. We tried using the rworldmap for the heat map, but it was not being knit into the output of the html file. For generating this heat map, we have used maptools library and given each region a color based on its case count. The issue with this map is that the legend is not visible. One cannot know the exact count of the region count by the map, but can compare the count of victims for all countries.

```
data1 <- read.csv("./data/The Global Dataset 3 Sept 2018.csv", na.strings = "-99")
drop <- c('terms.use')
data1 <- data1[, !(names(data1) %in% drop)]

countries <- read.csv('./data/all.csv')

data1 <- merge(x = data1,
              y = countries[,c('alpha.2', 'region', 'sub.region')],
              by.x = "citizenship",
              by.y = "alpha.2",
              all.x = TRUE)

p <- ggplot() + geom_bar(data = data1, aes(sub.region)) + coord_flip() +
     ggtitle("Count of victims based on their nationality")

ggplotly(p)
```

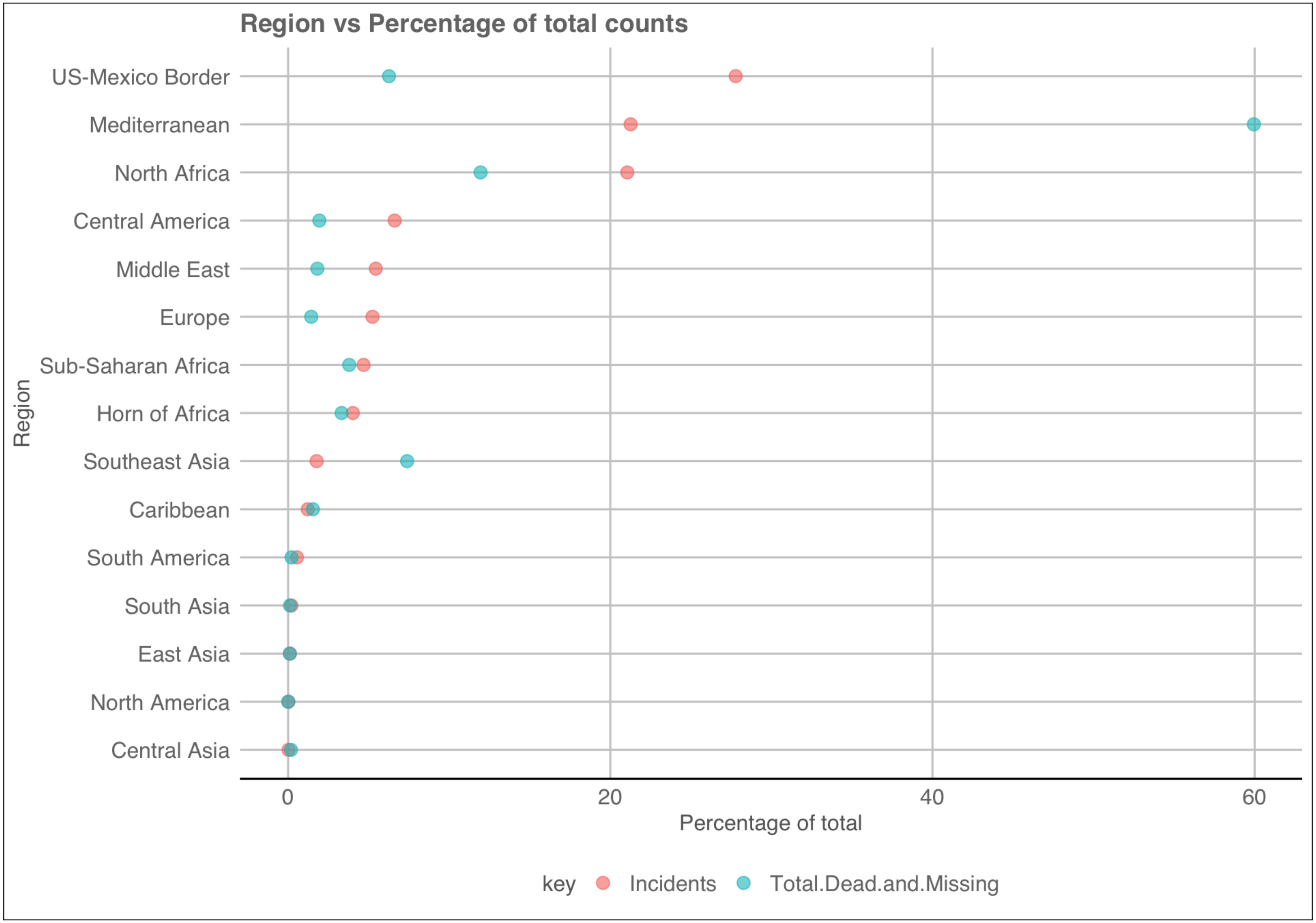


This plot shows the victims based on their country of citizenship.

4.2 The Missing Migrants dataset

Let us have a look at the Regions where the most incidents take place and where there are the most fatalities,

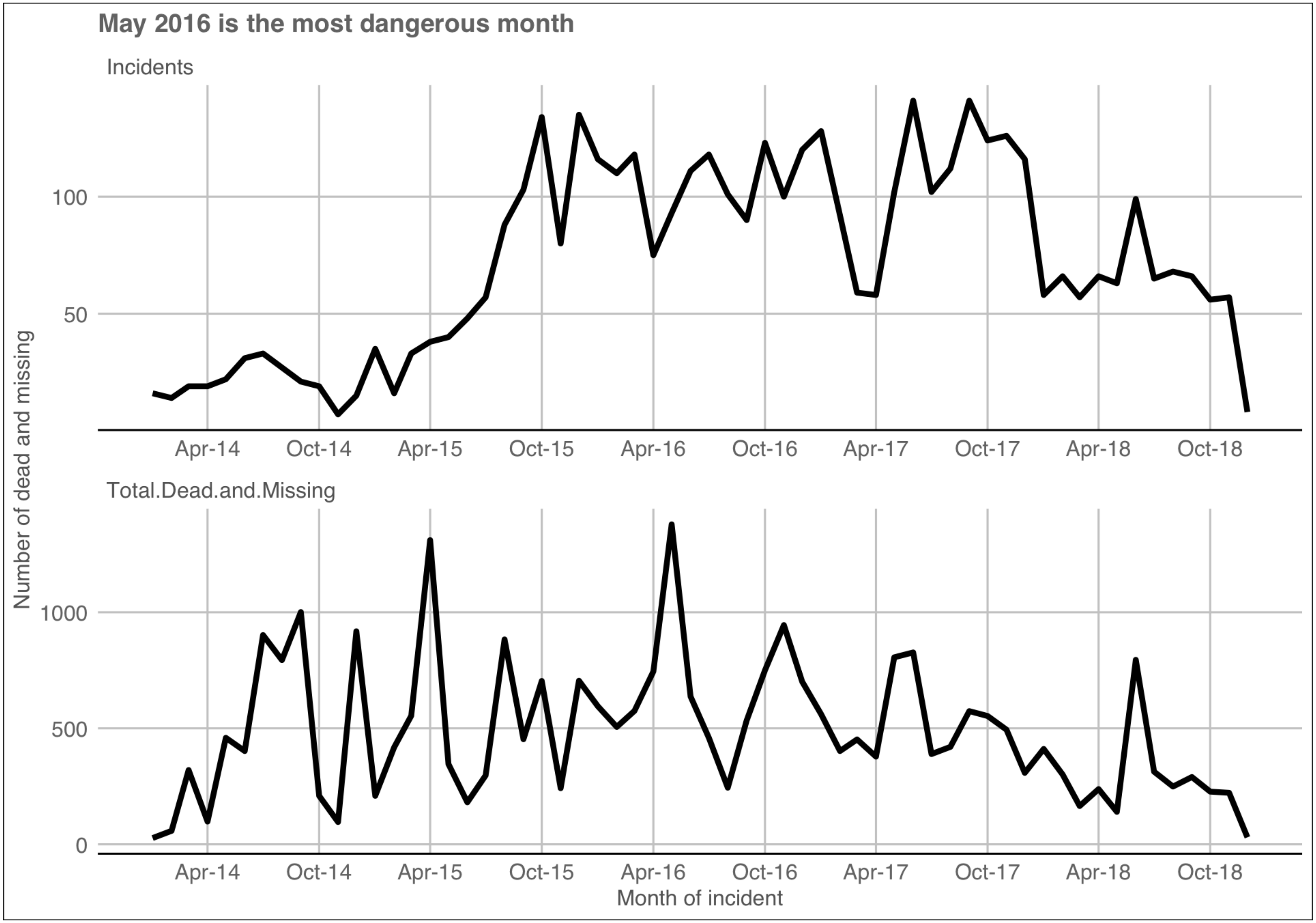
```
data$Incidents <- 1
small_data <- aggregate(Incidents ~ Region.of.Incident, data, sum)
small_data$Incidents <- (small_data$Incidents/ sum(small_data$Incidents)) * 100

small_data_2 <- aggregate(Total.Dead.and.Missing ~ Region.of.Incident, data, sum)
small_data_2$Total.Dead.and.Missing <- (small_data_2$Total.Dead.and.Missing/ sum(small_data_2$Total.Dead.and.Miss
ing)) * 100

small_data$Region.of.Incident <- reorder(small_data$Region.of.Incident, small_data$Incidents)
small_data <- merge(small_data, small_data_2)

small_data <- gather(small_data, ...= -Region.of.Incident)

ggplot() +
  geom_point(data = small_data, aes(x = Region.of.Incident, y = value, color = key), alpha = 0.6, size = 3) +
  coord_flip() +
  theme_gdocs() +
  xlab("Region") +
  ylab("Percentage of total") +
  ggtitle("Region vs Percentage of total counts") +
  theme(plot.title = element_text(size = 14, face = "bold")) +
  theme(legend.position="bottom",legend.direction="horizontal")
```



As can be seen, even though the US-Mexico border has the most number of incidents, that border does not have the most fatalities. Instead, the Mediterranean region has the most fatalities. It is also not far behind in terms of the number of incidents. An even more surprising part is that Southeast Asia does not have that many migrant incidents, but a lot more fatalities with respect to the migrant incidents.

Let us have a look at the trend of the number of dead and missing per time,

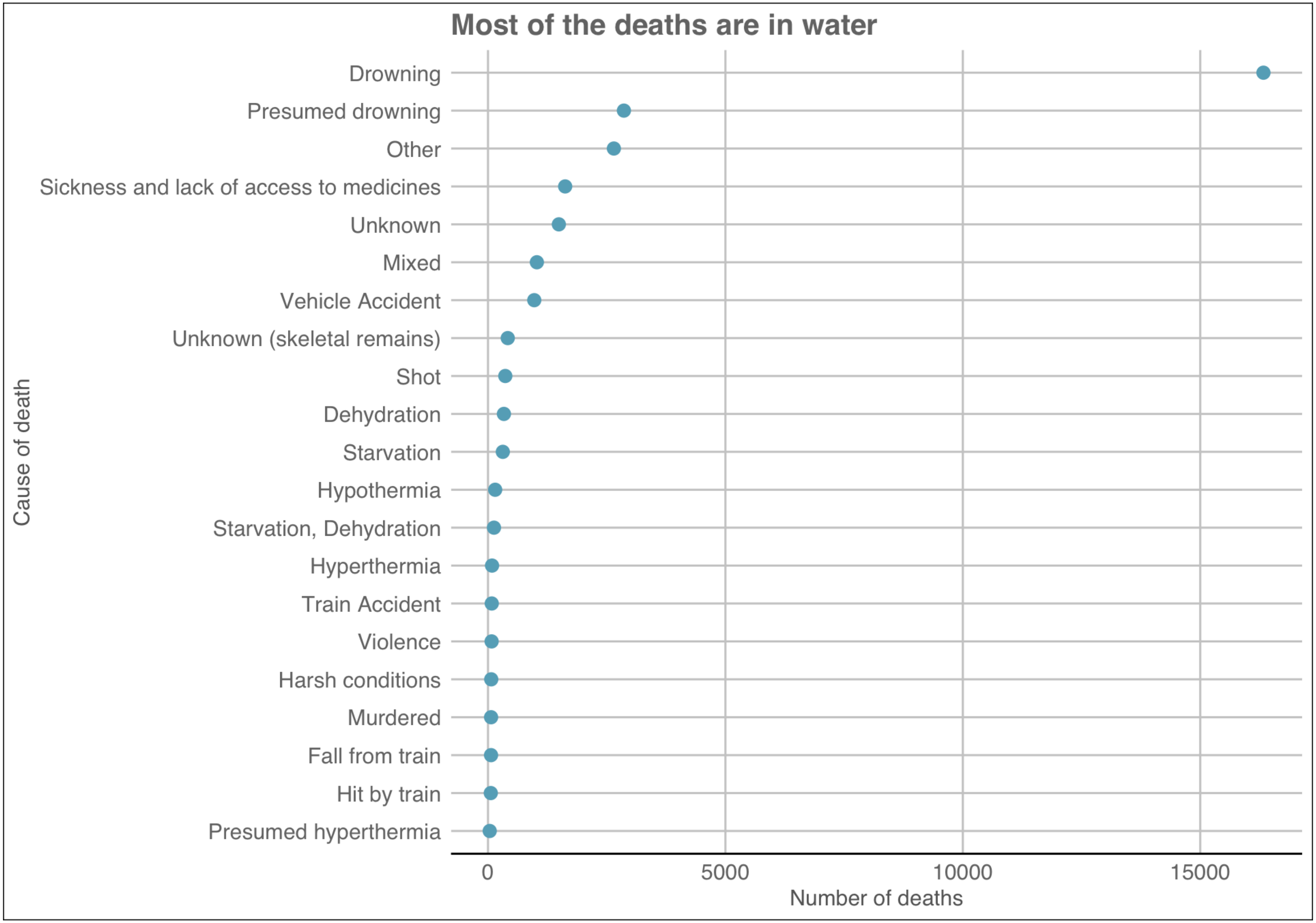
```
data$corrected_dates <- paste0(format(mdy(data$Reported.Date), format="%y-%m"),"-","01")
date_deaths_data <- aggregate(Total.Dead.and.Missing ~ corrected_dates, data, sum)
date_counts_data <- aggregate(Incidents ~ corrected_dates, data, sum)
date_deaths_data <- merge(date_deaths_data, date_counts_data)
date_deaths_data <- gather(date_deaths_data, ...=-corrected_dates)

# date_deaths_data$Total.Dead.and.Missing.3 <- SMA(date_deaths_data$Total.Dead.and.Missing,n=3)
#ggplot() +
#  geom_smooth(data = date_deaths_data, aes(x = corrected_dates, y = Total.Dead.and.Missing), method = "loess", f
ormula = y ~ x, size = # 1) +
#  scale_x_date(date_breaks = "6 months" , date_labels = "%b-%y")

ggplot() +
    geom_line(data = date_deaths_data, aes(x = ymd(corrected_dates), y = value), color = "black",size = 1.5) +
    scale_x_date(date_breaks = "6 months" , date_labels = "%b-%y") +
    theme_gdocs() +
  xlab("Month of incident") +
  ylab("Number of dead and missing") +
  ggtitle("May 2016 is the most dangerous month") +
  theme(plot.title = element_text(size = 14, face = "bold")) +
  facet_wrap( ~ key, ncol = 1, scales = "free")
```



As can be seen, there are huge peaks in the number of fatalities in June 2016 and April 2015, but the number of incidents do not correspond to the same peaks, which may show that there were some incidents that took place that month which had a lot of people involved. Also, another promising fact is that as time goes on, we can see that the number of fatalities are decreasing over time, and so are the number of incidents.

Let us plot which causes of death lead to the most fatalities,

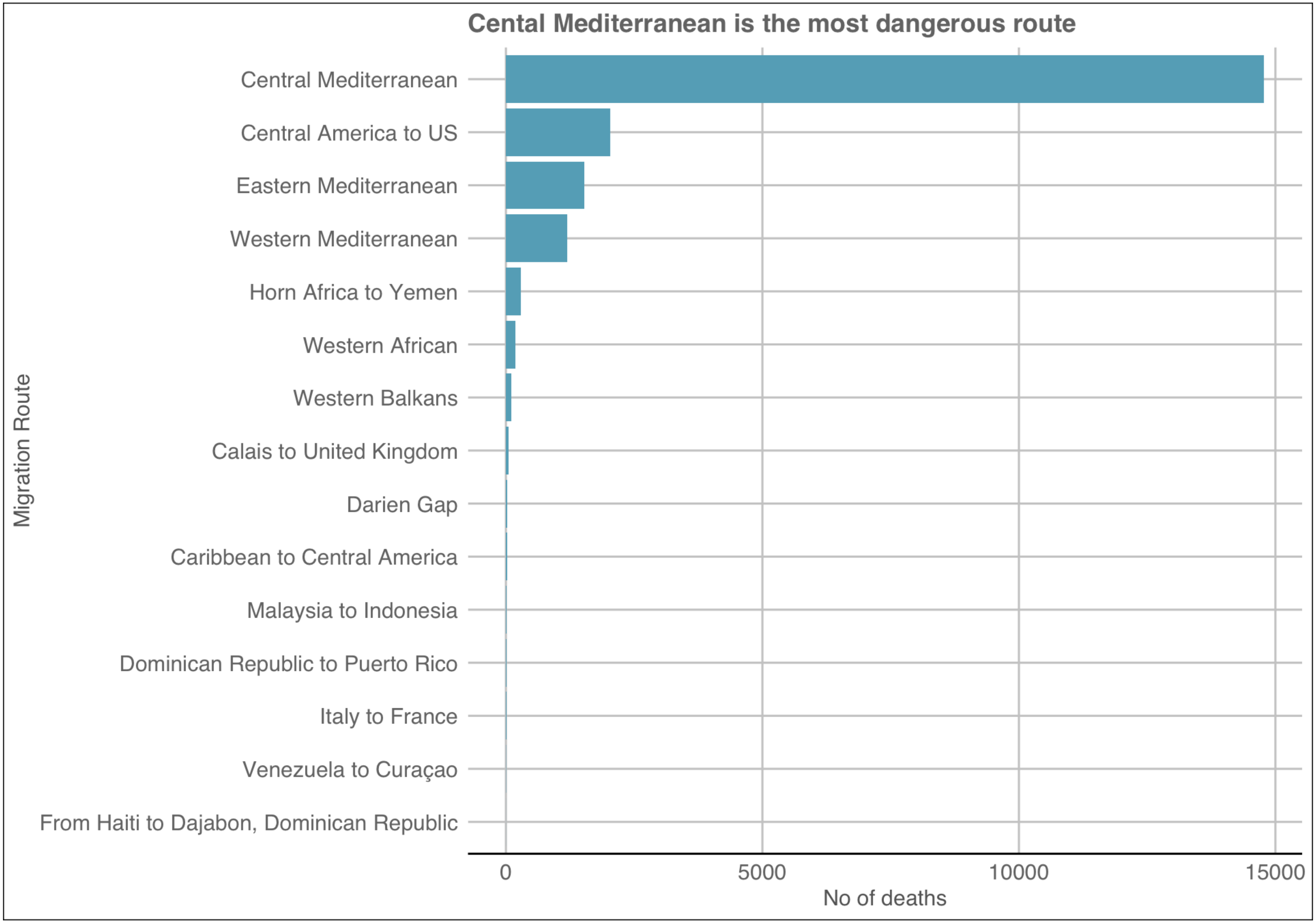
```
n <- 20
fct_causes <- names(sort(table(data$Cause.of.Death), decreasing = TRUE)[1:n])
data$count <- 1
data <- mutate(data, Cause.of.Death = fct_other(Cause.of.Death, keep = fct_causes, other_level = 'Other'))
data$Cause.of.Death <- fct_infreq(data$Cause.of.Death)
data$Cause.of.Death <- fct_relevel(factor(data$Cause.of.Death), "Other", after = Inf)
small_data <- aggregate(Total.Dead.and.Missing ~ Cause.of.Death, data, sum)
small_data$Cause.of.Death <- reorder(small_data$Cause.of.Death, small_data$Total.Dead.and.Missing)

ggplot() +
  geom_point(data = small_data, aes(x = Cause.of.Death, y = Total.Dead.and.Missing), size = 3) +
  coord_flip() +
  ylab("Number of deaths") +
  xlab("Cause of death") +
  ggtitle("Most of the deaths are in water") +
  theme_gdocs() +
  theme(plot.title = element_text(size = 16, face = "bold"))
```



It can be seen that the most fatal deaths are on the sea, due to drowning, followed by sicknesses, which would make sense as most journeys are long and a lot of times the migrants do not have access to the minimum medicines and resources.

We can also observe the most dangerous migration routes, ie, the routes which cause the most fatalities,

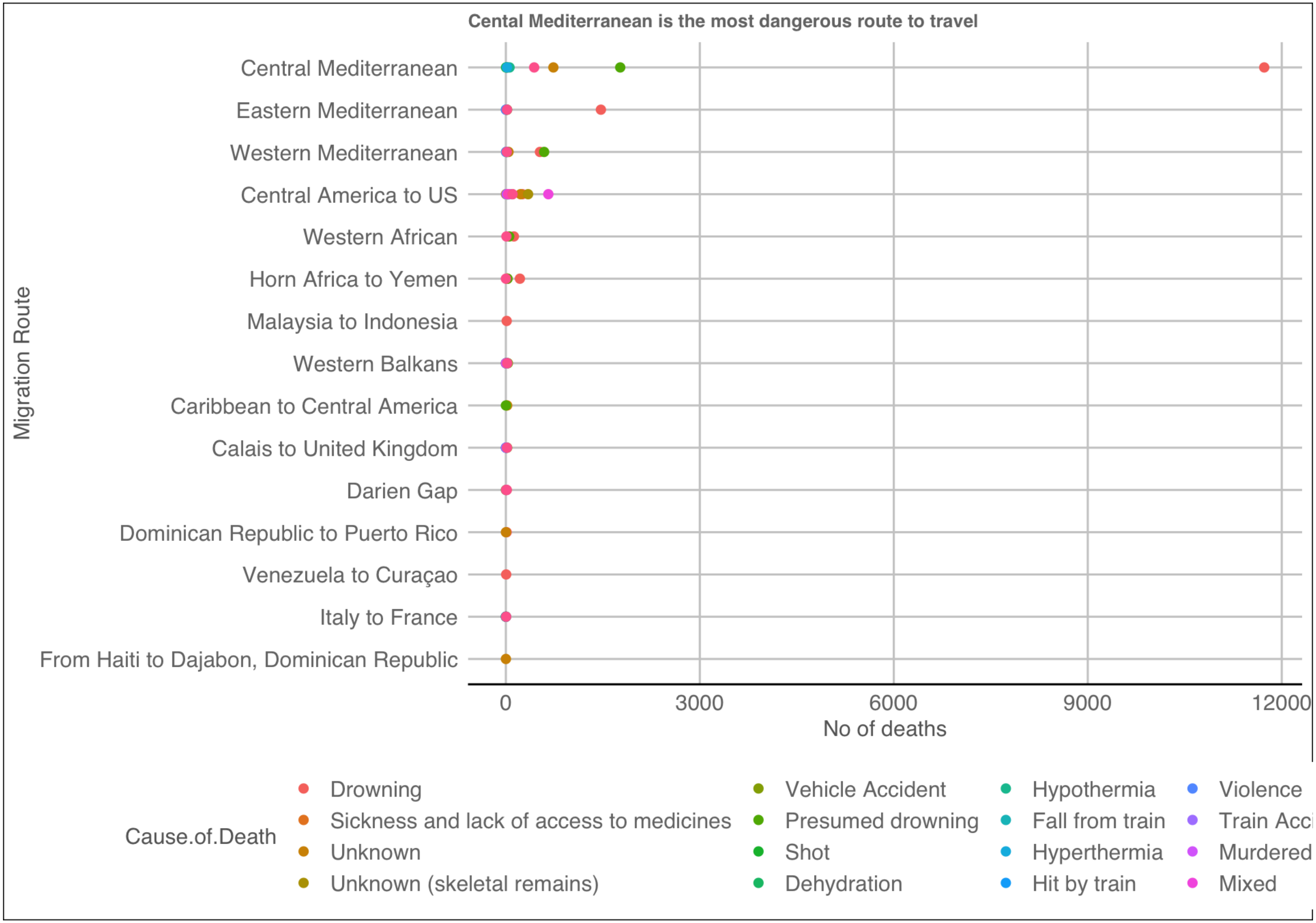
```
small_data <- aggregate(Total.Dead.and.Missing ~Migration.Route, data, sum)

ggplot() +
  geom_bar(data = small_data, aes(x = reorder(Migration.Route, Total.Dead.and.Missing), y = Total.Dead.and.Missin
g), stat = "identity") +
  theme_gdocs() +
  xlab("Migration Route") +
  ylab("No of deaths") +
  coord_flip() +
  ggtitle("Cental Mediterranean is the most dangerous route") +
  theme(plot.title = element_text(size = 14, face = "bold"))
```



As can be seen, the most fatal migration route is the Central Mediterranean, followed by route from the Central America to the US. The numbers might be population biased, as a huge number of people migrate along that route. Also, due to the rough Mediterranean Sea, it would be dangerous to cross over from Africa to Europe.

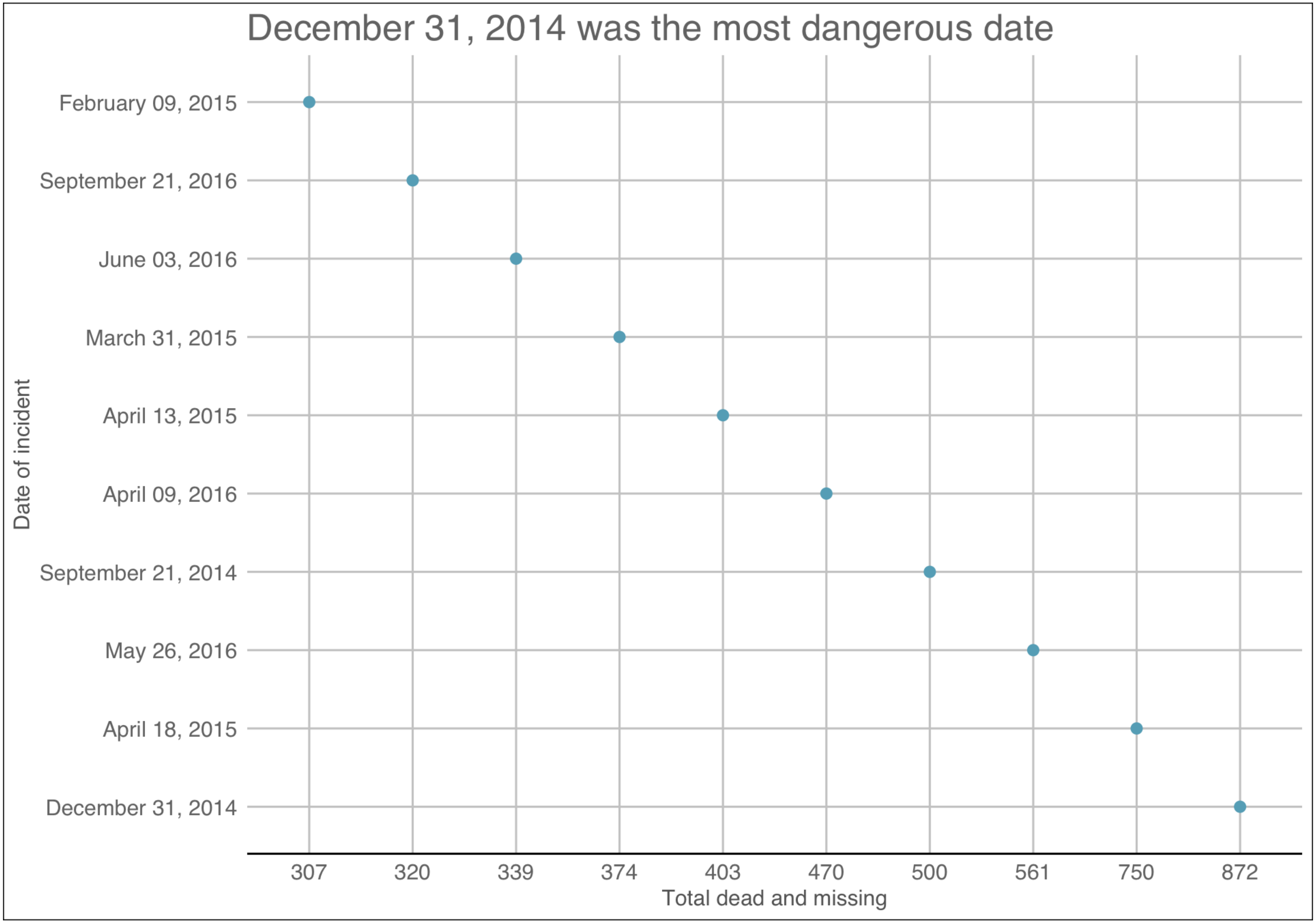
The above hypothesis can be corroborated by the below graph, where we can see that the most dangerous migration route is the Central Mediterranean, and the biggest cause of death along that route is Drowning,

```
small_data <- aggregate(Total.Dead.and.Missing ~Migration.Route + Cause.of.Death, data, sum)

ggplot() +
  geom_point(data = small_data, aes(x = reorder(Migration.Route, Total.Dead.and.Missing), y = Total.Dead.and.Miss
ing, color = Cause.of.Death), stat = "identity", size = 2) +
  theme_gdocs() +
  xlab("Migration Route") +
  ylab("No of deaths") +
  coord_flip() +
  ggtitle("Cental Mediterranean is the most dangerous route to travel") +
  theme(plot.title = element_text(size = 10, face = "bold")) +
  theme(legend.position="bottom",legend.direction="horizontal")
```



Most dangerous dates,

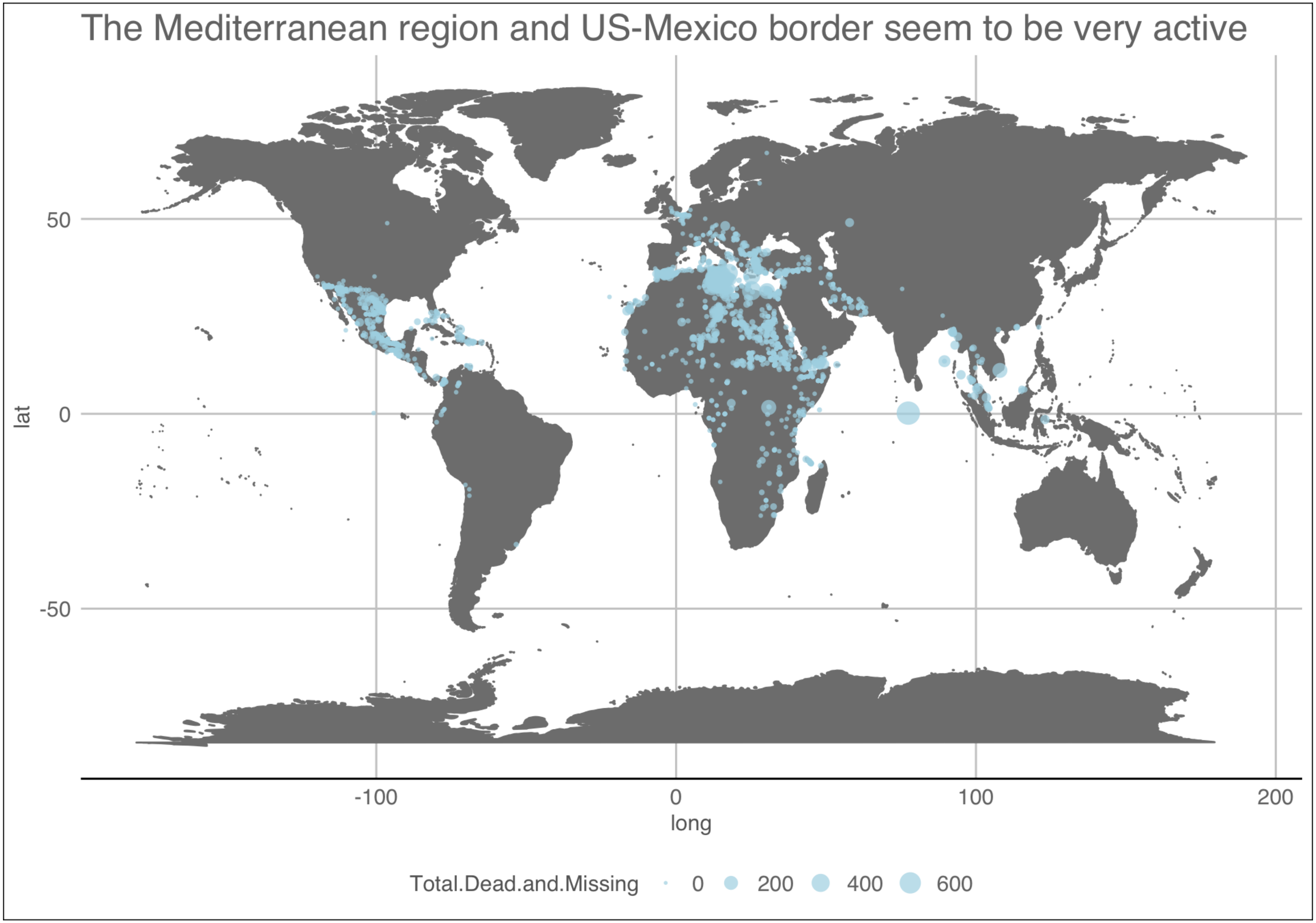
```
dang_dates <- aggregate(Total.Dead.and.Missing ~ Reported.Date, data, sum)
dang_dates <- dang_dates[order(dang_dates$Total.Dead.and.Missing, decreasing = TRUE),]
dang_dates$Reported.Date <- factor(dang_dates$Reported.Date)
n <- 10
dang_dates <- top_n(dang_dates, n = n, Total.Dead.and.Missing)
dang_dates$Total.Dead.and.Missing <- factor(dang_dates$Total.Dead.and.Missing)
dang_dates <- dang_dates[order(dang_dates$Total.Dead.and.Missing, decreasing = TRUE),]
dang_dates$Reported.Date <- factor(dang_dates$Reported.Date, levels= dang_dates[order(dang_dates$Total.Dead.and.M
issing, decreasing = TRUE),'Reported.Date'], ordered = TRUE)

ggplot() +
  geom_point(data = dang_dates, aes(x = Reported.Date, y = Total.Dead.and.Missing), stat = "identity", size = 2.5
) +
  coord_flip() +
  ylab("Total dead and missing") +
  xlab("Date of incident") +
  ggtitle("December 31, 2014 was the most dangerous date") +
  theme_gdocs()
```



It seems that the most dangerous date was December 31st, 2014, followed by April 18th, 2015. The first date corresponds with the information of "ghost" ships, which were abandoned by the crew, and set sail for Europe. The second incident corresponds to a news of approximately 700 migrants who died in a Mediterranean shipwreak.

We can plot these areas on the world map to visually look the hotspots of migrant incidents,

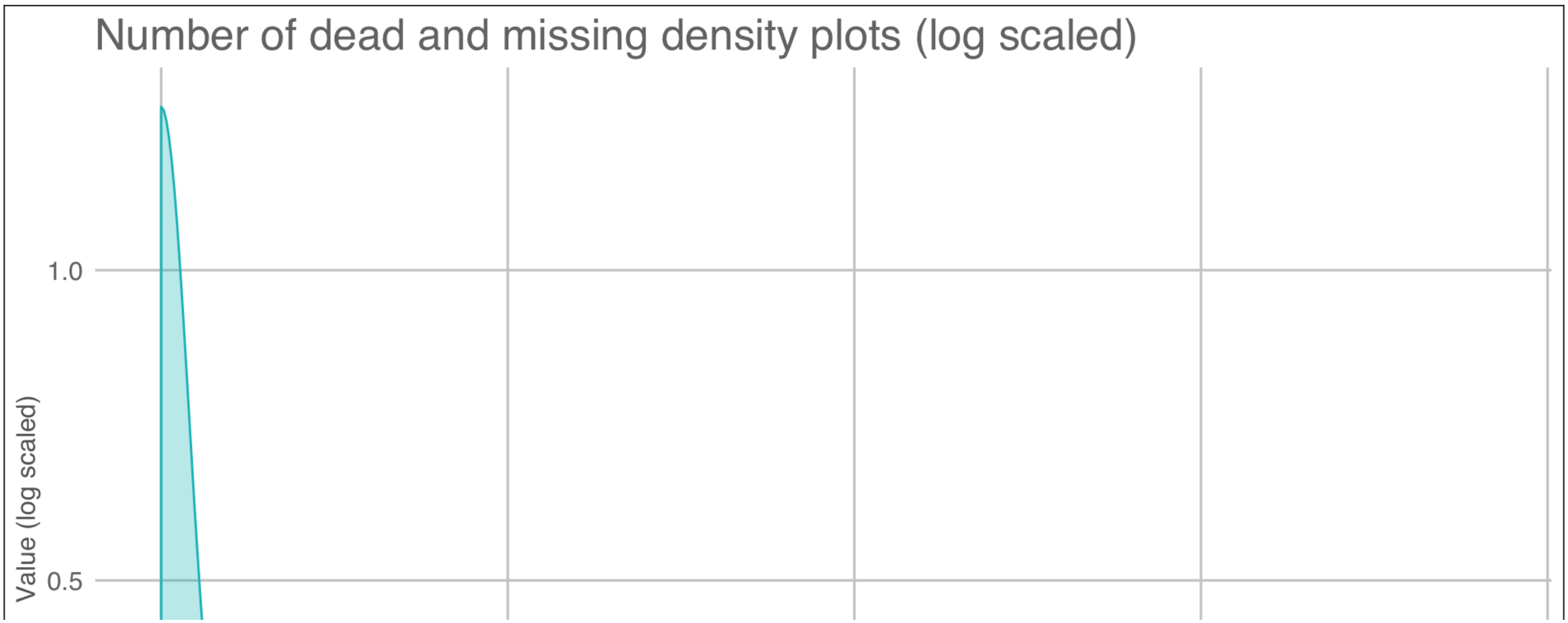
```
data <- separate(data = data, col = Location.Coordinates, into = c("coord.y", "coord.x"), sep = ",")
data <- data[(as.numeric(data$coord.x) >= -180) & (as.numeric(data$coord.x) <= 180) & (as.numeric(data$coord.y) >
= -90) & (as.numeric(data$coord.y) <= 90),]

mp <- NULL
mapWorld <- borders("world", colour="gray50", fill="gray50")
ggplot() +
  mapWorld +
  geom_point(data = data, aes(x = as.numeric(coord.x), y = as.numeric(coord.y), size = Total.Dead.and.Missing), c
olor = 'light blue', fill = 'black', stroke = 0, alpha = 0.7) +
  ggtitle("The Mediterranean region and US-Mexico border seem to be very active") +
  theme_gdocs() +
  theme(legend.position="bottom",legend.direction="horizontal")
```

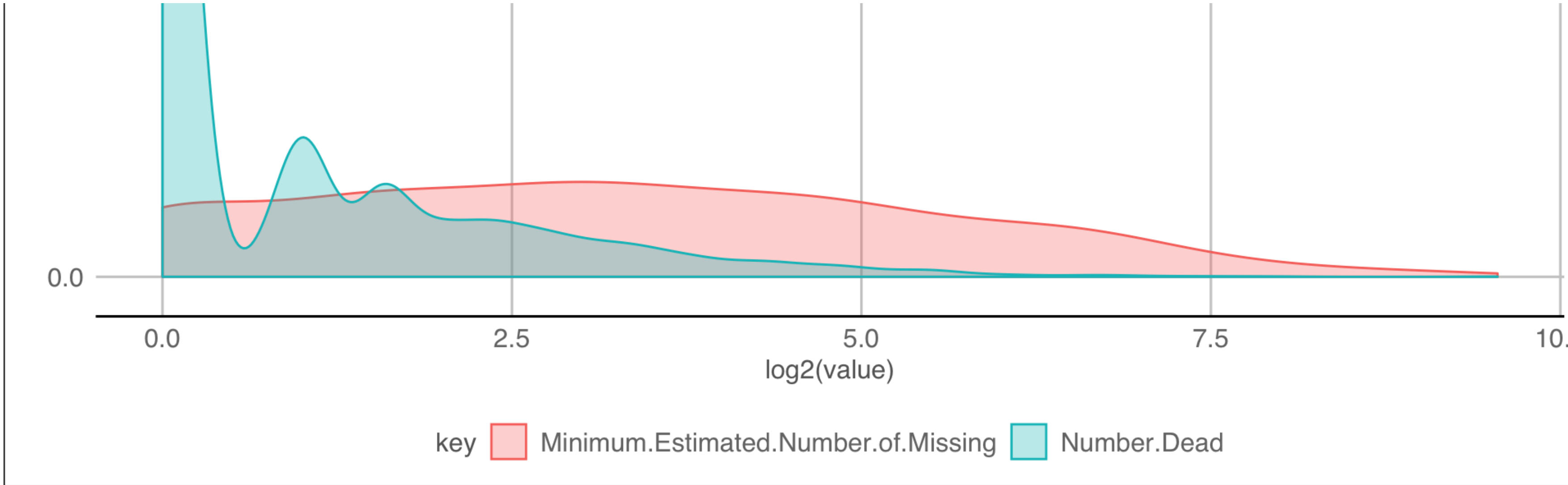


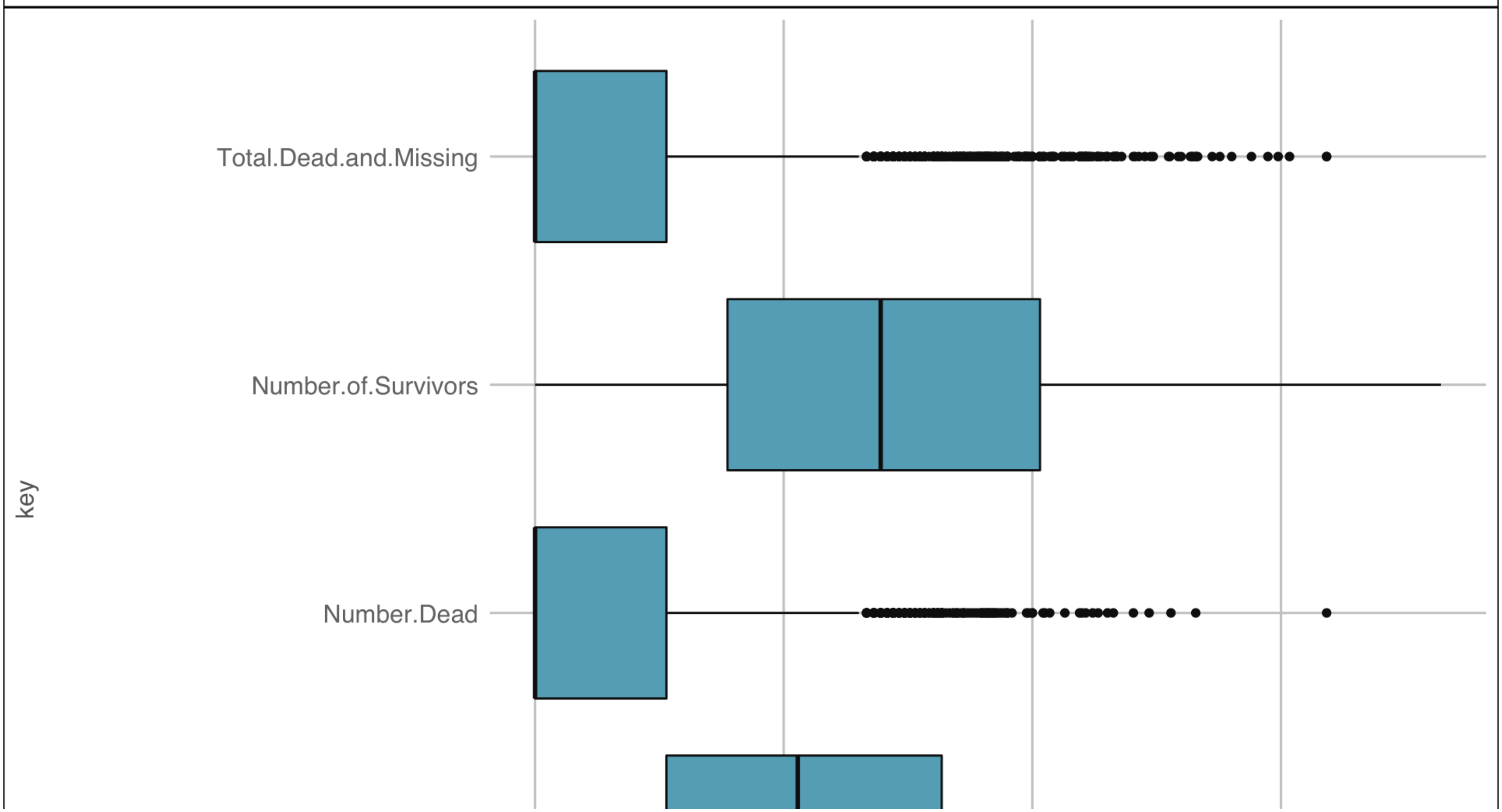
As can be seen, the US-Mexico border and the Mediterranean area is the most active area in terms of migrant incidents.

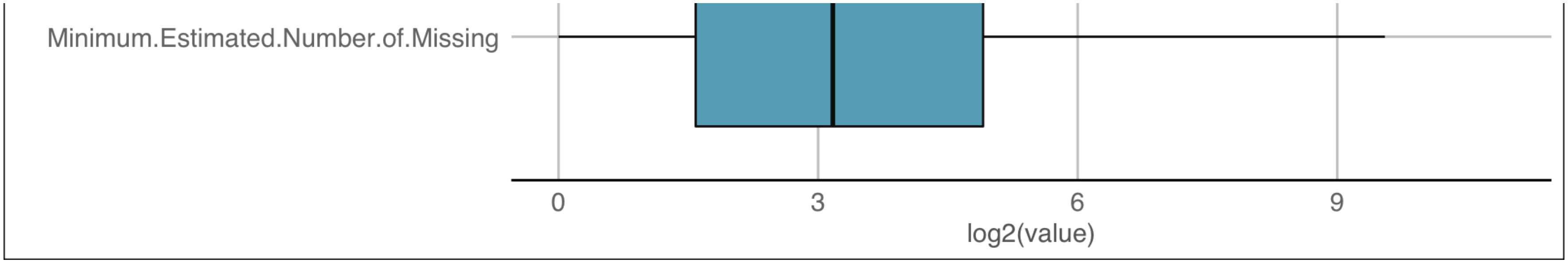
We can also look at the trends of dead and missing by density and the boxplot of the number of dead, the estimated missing people, the dead people and the number of survivors,

```
gathered_data <- gather(data[,c('Number.Dead', 'Minimum.Estimated.Number.of.Missing')])

p1 <- ggplot() +
  geom_density(data = gathered_data, aes(log2(value), fill = key, color = key), alpha = 0.3) +
  ggtitle("Number of dead and missing density plots (log scaled)") +
  theme_gdocs() +
  theme(legend.position="bottom",legend.direction="horizontal") +
  ylab("Value (log scaled)")

##Boxplots of missing and dead,
cols <- c('Web.ID','Number.Dead', 'Minimum.Estimated.Number.of.Missing', 'Total.Dead.and.Missing', 'Number.of.Sur
vivors')
melted_data <- gather(data = data[, names(data) %in% cols], .. = -Web.ID)


p2 <- ggplot() +
  geom_boxplot(data = melted_data, aes(x = key, y = log2(value)), aes = 0.8) +
  coord_flip() +
  theme_gdocs()

grid.arrange(p1,p2, ncol = 1)
```




0.0
0.0
2.5
5.0
7.5
10.0
log2(value)
key
Minimum.Estimated.Number.of.Missing
Number.Dead

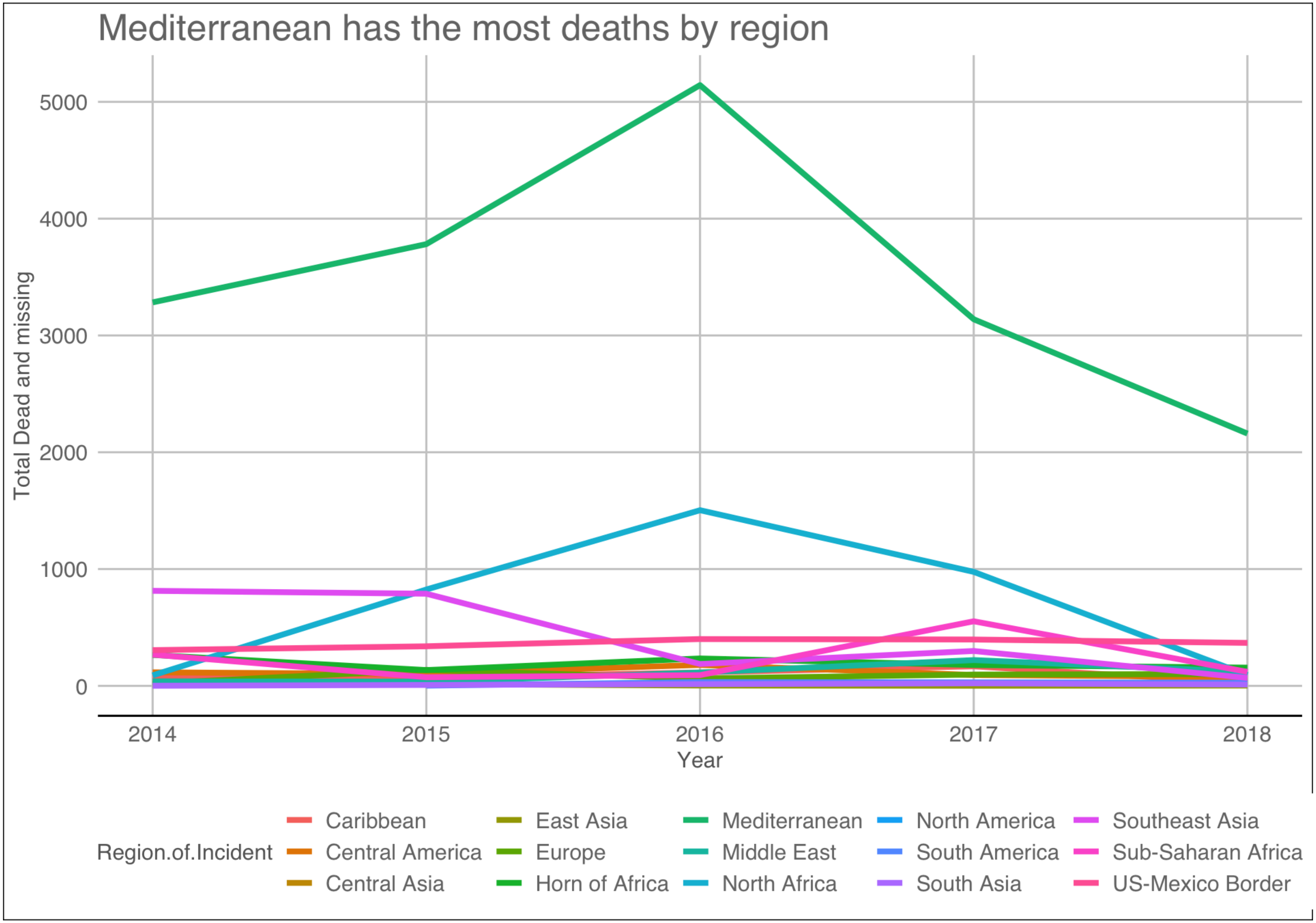

Total.Dead.and.Missing
Number.of.Survivors
Number.Dead
key



As can be seen, the number of dead has mostly small values.

Plot number of deaths by region,

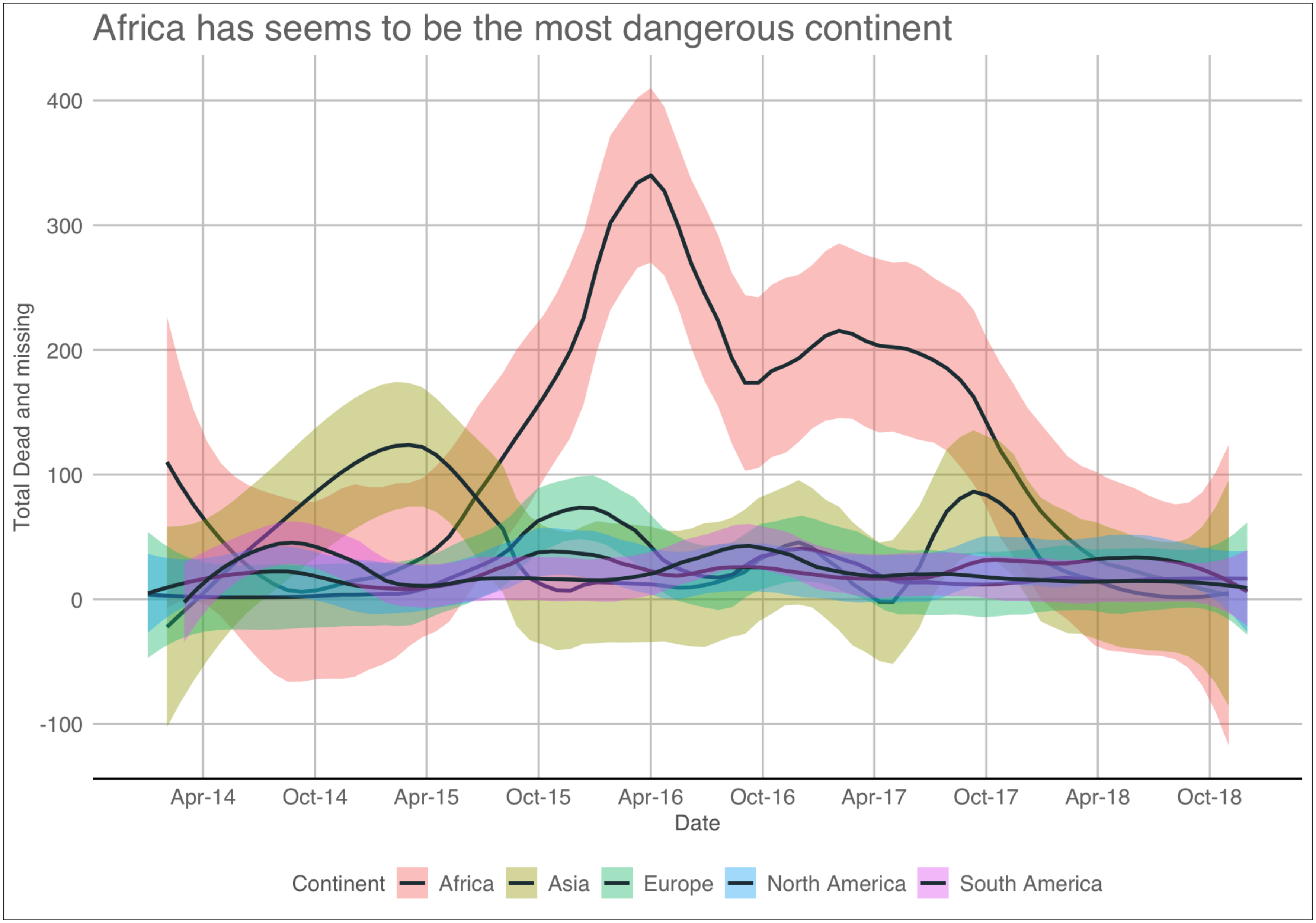
```
date_deaths_data <- aggregate(Total.Dead.and.Missing ~ Reported.Year + Region.of.Incident, data, sum)

ggplot() +
  geom_line(data = date_deaths_data, aes(x = Reported.Year, y = Total.Dead.and.Missing, color = Region.of.Inciden
t),size = 1.5) +
  theme_gdocs() +
  theme(legend.position="bottom",legend.direction="horizontal") +
  ggtitle("Mediterranean has the most deaths by region") +
  xlab("Year") +
  ylab("Total Dead and missing") +
  guides(fill=guide_legend(title="Region of Incident"))
```



We can see that the Most dead and missing are in the region Mediterranean, but the trend is decreasing in nature.

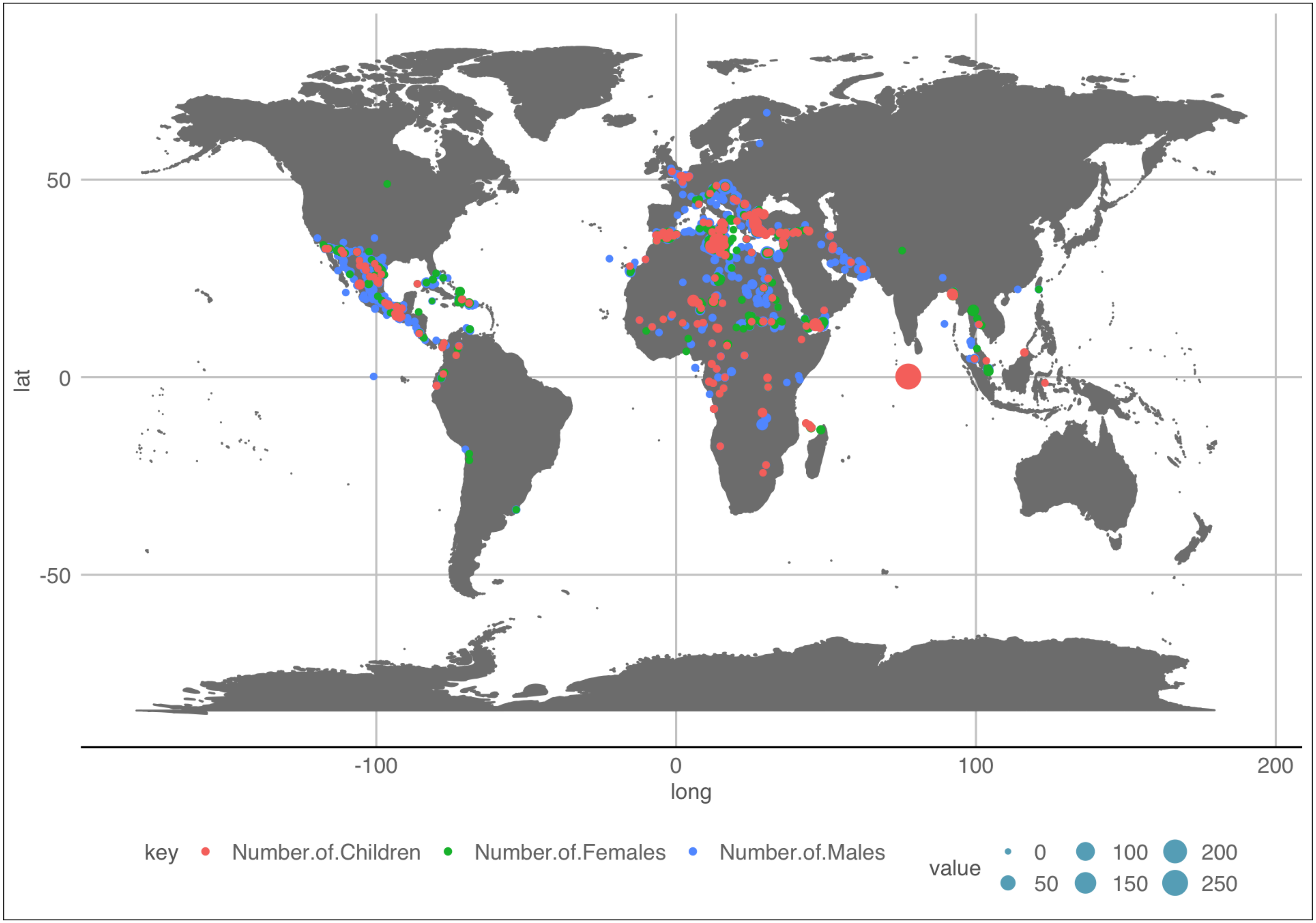
This can be coorborated by the below graph on a continent level,

```
coords2continent = function(points)
{
  countriesSP <- getMap(resolution='high')
  pointsSP = SpatialPoints(points, proj4string=CRS(proj4string(countriesSP)))
  indices = over(pointsSP, countriesSP)
  indices$REGION   # returns the continent (7 continent model)
}

points <- data.frame(lon=sign(as.numeric(data$coord.x)) * ceiling(abs(as.numeric(data$coord.x)) * 100) / 100,
                     lat =sign(as.numeric(data$coord.y)) * ceiling(abs(as.numeric(data$coord.y)) * 100) / 100)

continents <- coords2continent(points)
data$continent <- continents

data$corrected_dates <- paste0(format(mdy(data$Reported.Date), format="%y-%m"),"-","01")
date_deaths_data <- aggregate(Total.Dead.and.Missing ~ corrected_dates + continent, data, sum)

ggplot() +
    geom_smooth(data = date_deaths_data, aes(x = ymd(corrected_dates), y = Total.Dead.and.Missing, fill = contine
nt), span = 0.3) +
    scale_x_date(date_breaks = "6 months" , date_labels = "%b-%y") +
  theme(legend.position="bottom") +
  theme(legend.position="bottom",legend.direction="horizontal") +
  ggtitle("Africa has seems to be the most dangerous continent") +
  xlab("Date") +
  ylab("Total Dead and missing") +
  guides(fill=guide_legend(title="Continent")) +
  theme_gdocs() +
  theme(legend.position="bottom",legend.direction="horizontal")
```



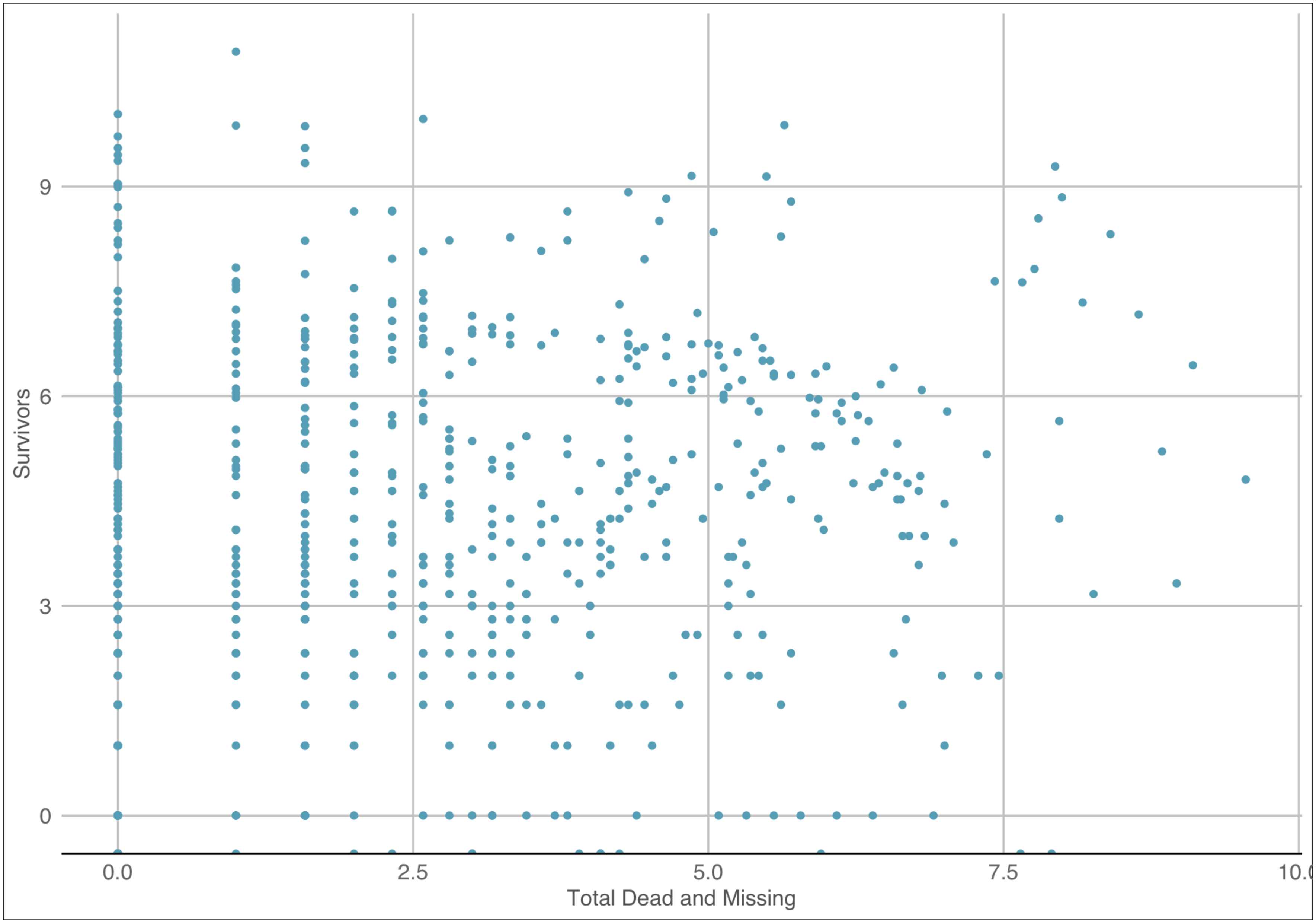
We can also look at the number of males, females and childern who died in the incidents,

```
cols <- c('Number.of.Males', 'Number.of.Females', 'Number.of.Children', 'coord.x', 'coord.y')
melted_data <- gather(data[,cols], ... = -c('coord.x', 'coord.y'))

mp <- NULL
mapWorld <- borders("world", colour="gray50", fill="gray50")
ggplot() +
  mapWorld +
  geom_point(data = melted_data, aes(x = as.numeric(coord.x), y = as.numeric(coord.y), size = value, color = ke
y), fill = 'light blue') +
  theme_gdocs() +
  theme(legend.position="bottom",legend.direction="horizontal")
```



We can look at a scatterplot of the Dead and missing vs the number of survivors,

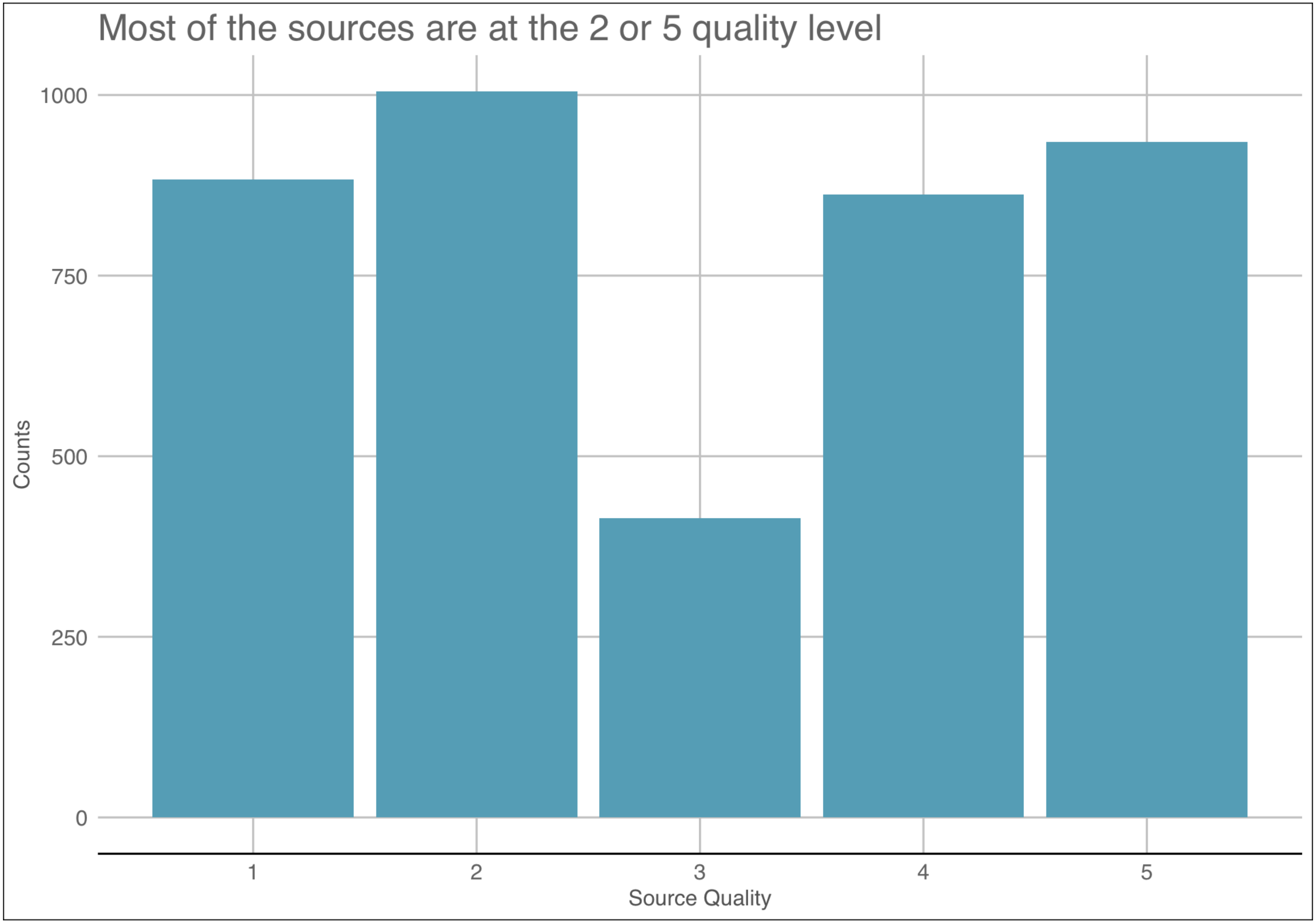
```
ggplot() +
  geom_point(data = data, aes(x = log2(Total.Dead.and.Missing), y = log2(Number.of.Survivors)), size = 1.5) +
  theme_gdocs() +
  xlab("Total Dead and Missing") +
  ylab("Survivors")
```



There does not seem to be any specific trend, which is a little surprising.

From the data source, the quality of source is a column which represents the number of source which cooroborate the incident,

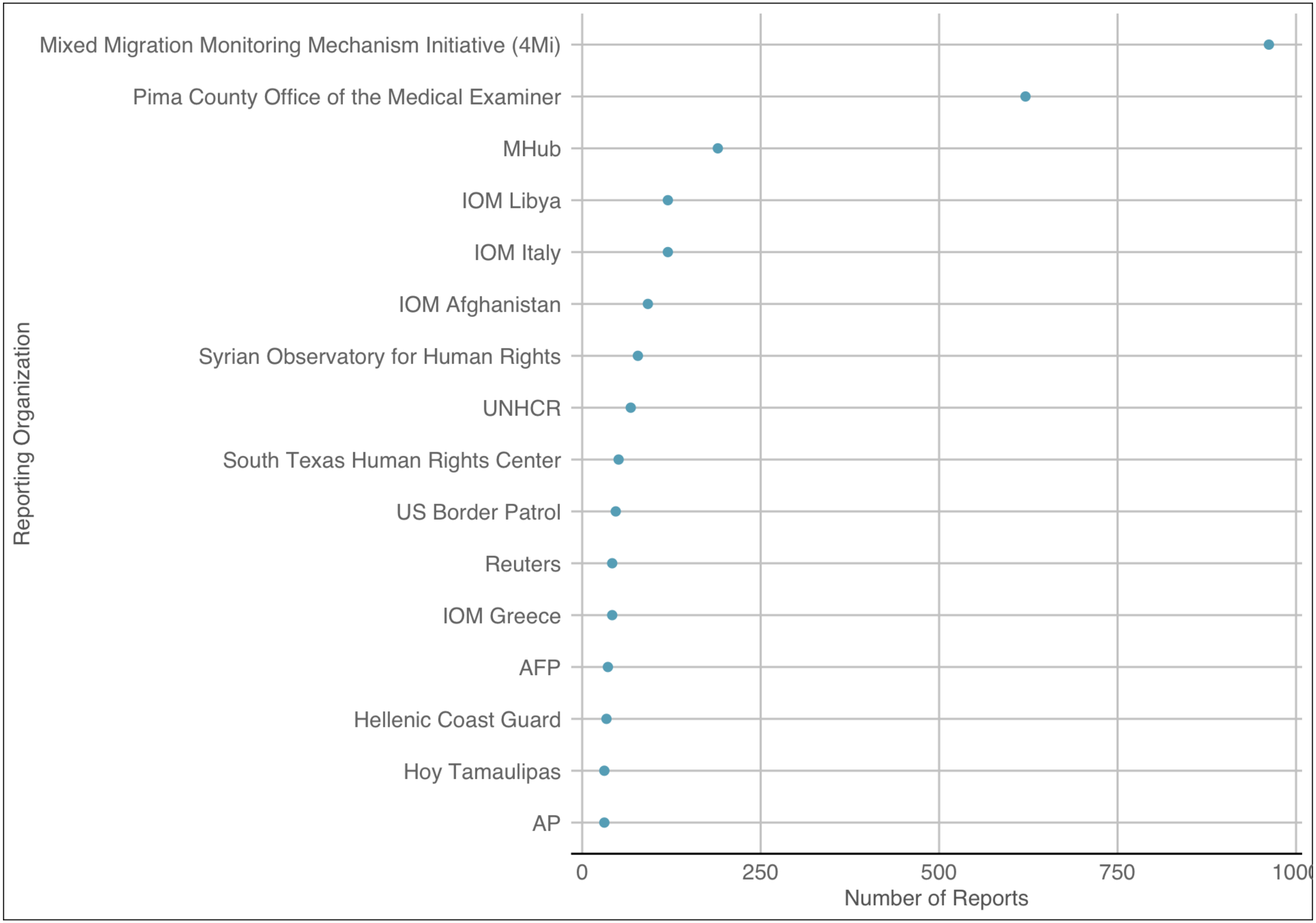
```
ggplot() +
  geom_bar(data = data, aes(Source.Quality)) +
  xlab("Source Quality") +
  ylab("Counts") +
  ggtitle("Most of the sources are at the 2 or 5 quality level") +
  theme_gdocs()
```



We can see that very few of our incidents are of 3-quality.

We can also look at the top news sources,

```
info_sources <- data %>% separate_rows(Information.Source, sep = ",") %>% select(c("Information.Source", "Source.
Quality"))
info_sources$count <- 1
count_sources <- aggregate(count ~ Information.Source, info_sources, sum)
n <- 15

count_sources <- top_n(count_sources, n=n, count)

ggplot(data = count_sources, aes(x=reorder(Information.Source, count), y=count))+
  geom_point(stat='identity', size = 2) +
  coord_flip() +
  theme_gdocs() +
  ylab("Number of Reports") +
  xlab("Reporting Organization")
```

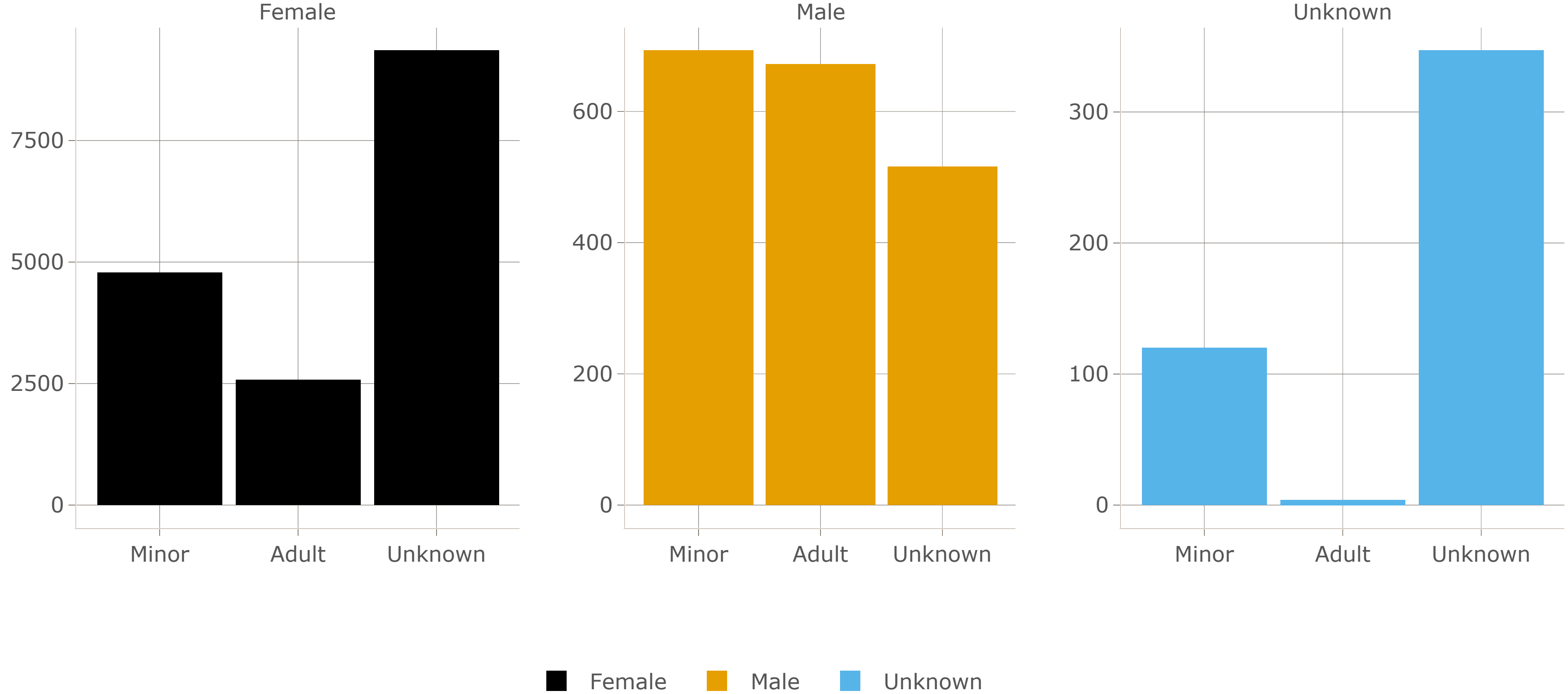


According to the above plot, the 4M Initiative and the Pima County Office are the biggest contributors to the incidents.

5. EXECUTIVE SUMMARY

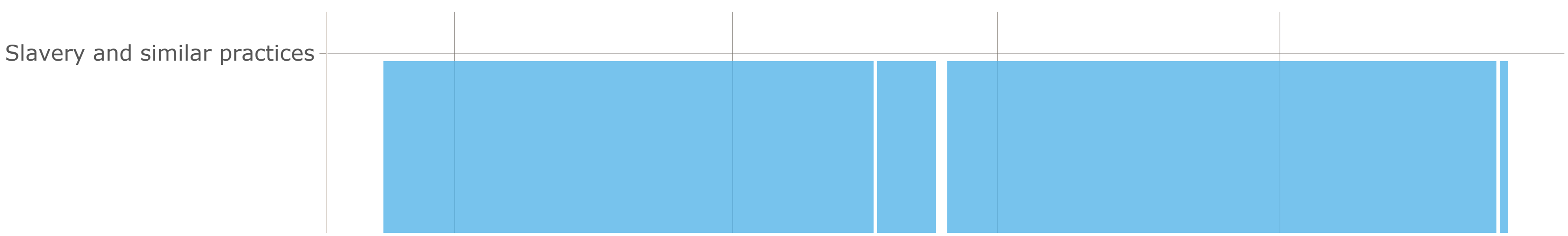
5.1 Human Trafficking Dataset

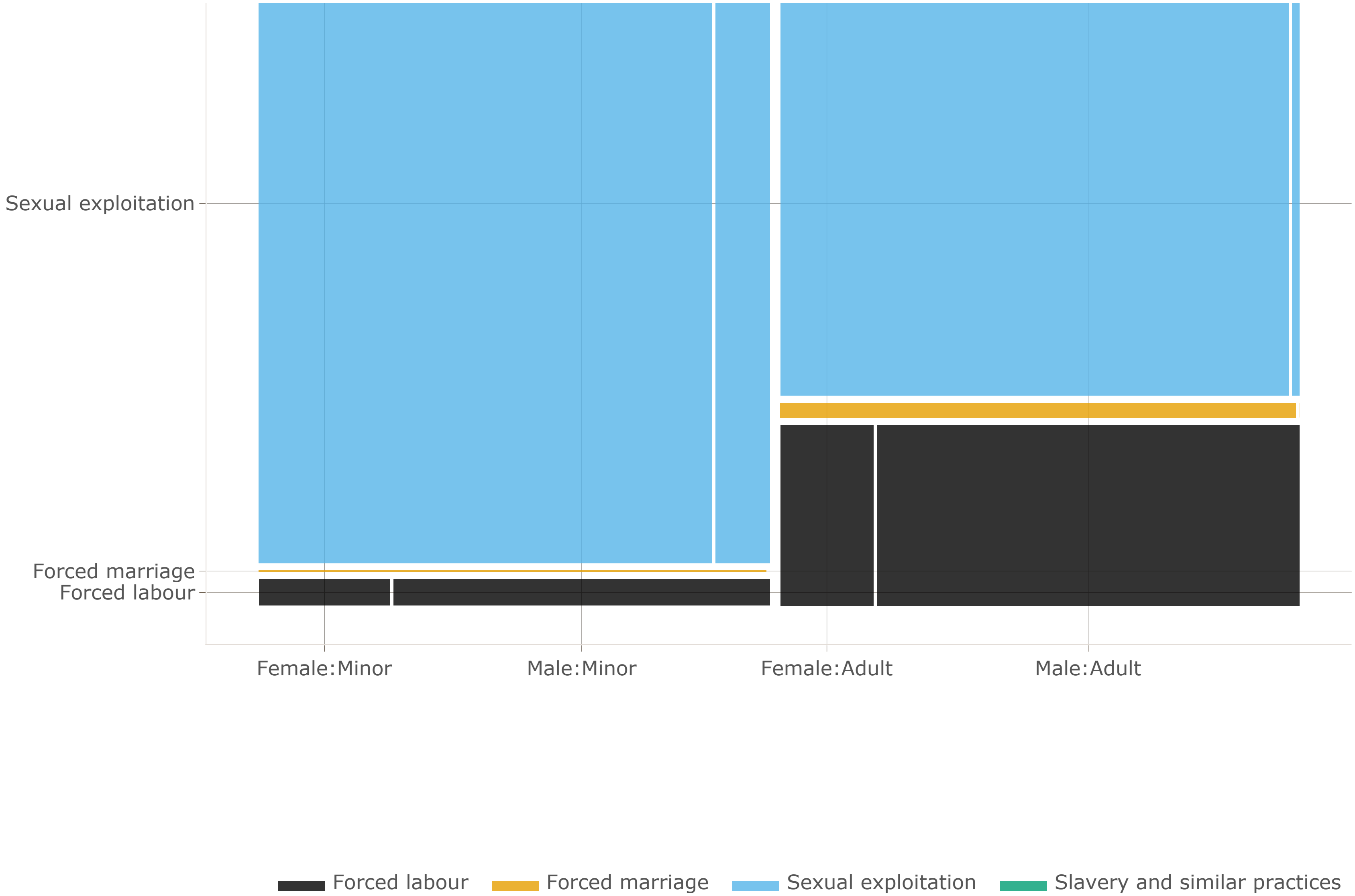
Count of People Trafficked by Gender



The graph shows the count of the people trafficked categorised by their gender. Each gender is then grouped by their Age. We can see that there are a lot of unkowns which means that the data has missing data which may be due to negligence while filling the survey forms or this data may be lost. We can see that the count of female victims that are trafficked is much larger than the number of male victims(consider the scales of each plot separately). Among females, minors,due to their innocence, are more susceptible to human trafficking than adults.

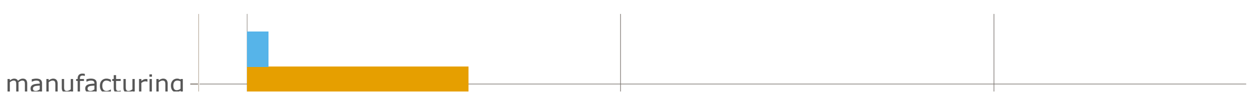
Type of Exploitation By Gender And Age

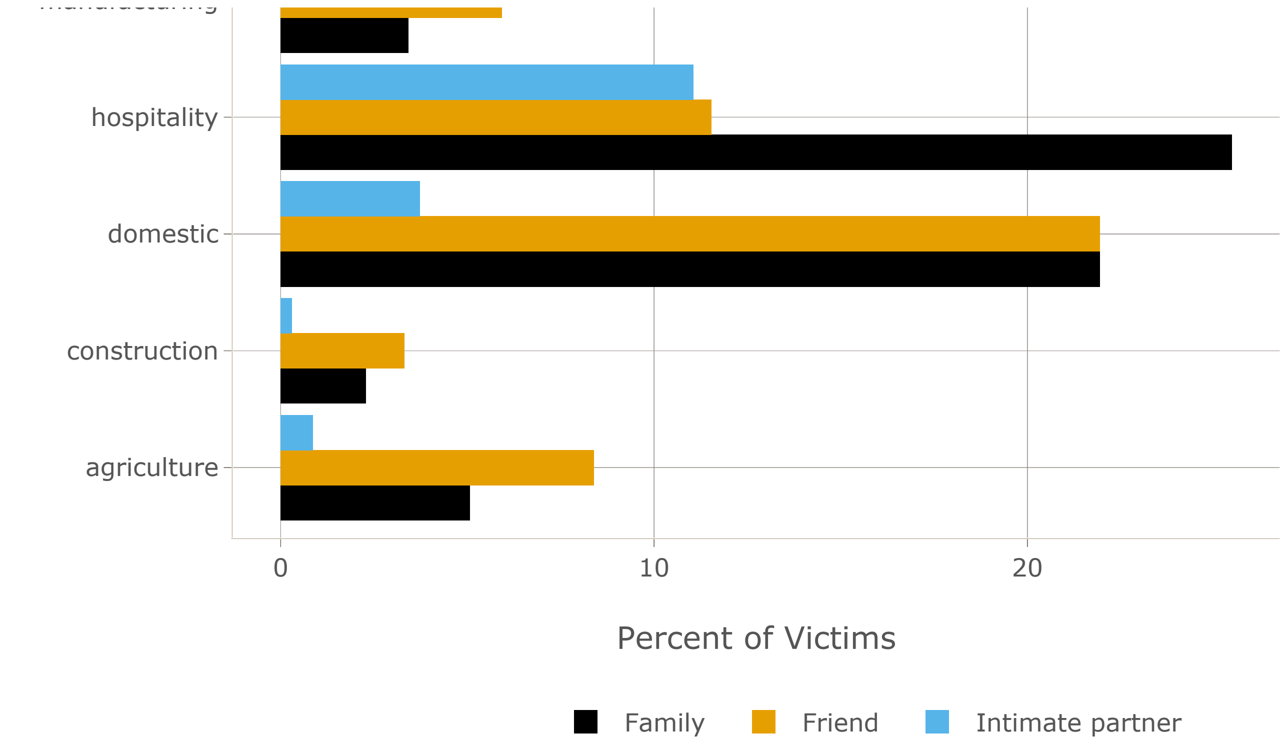




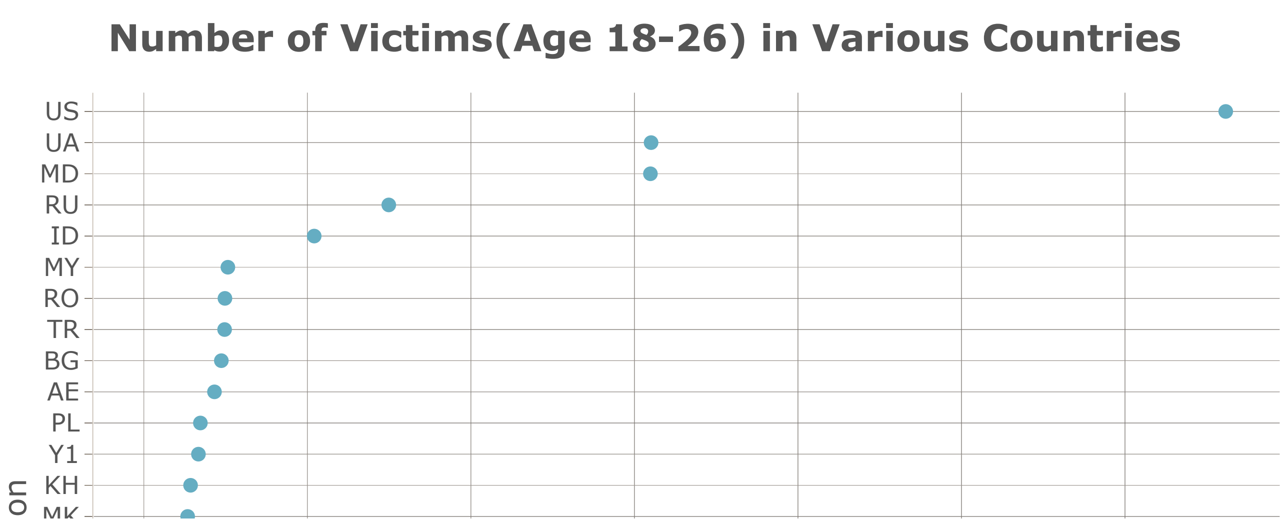
This plot shows the proportions of abuse performed on different classes of victims(Females and Males: Minors and Adults). It shows that sexual abuse is the most common form of abuse among females(as it has a larger area/width in the plot),whereas forced labour is the most common in males victims.

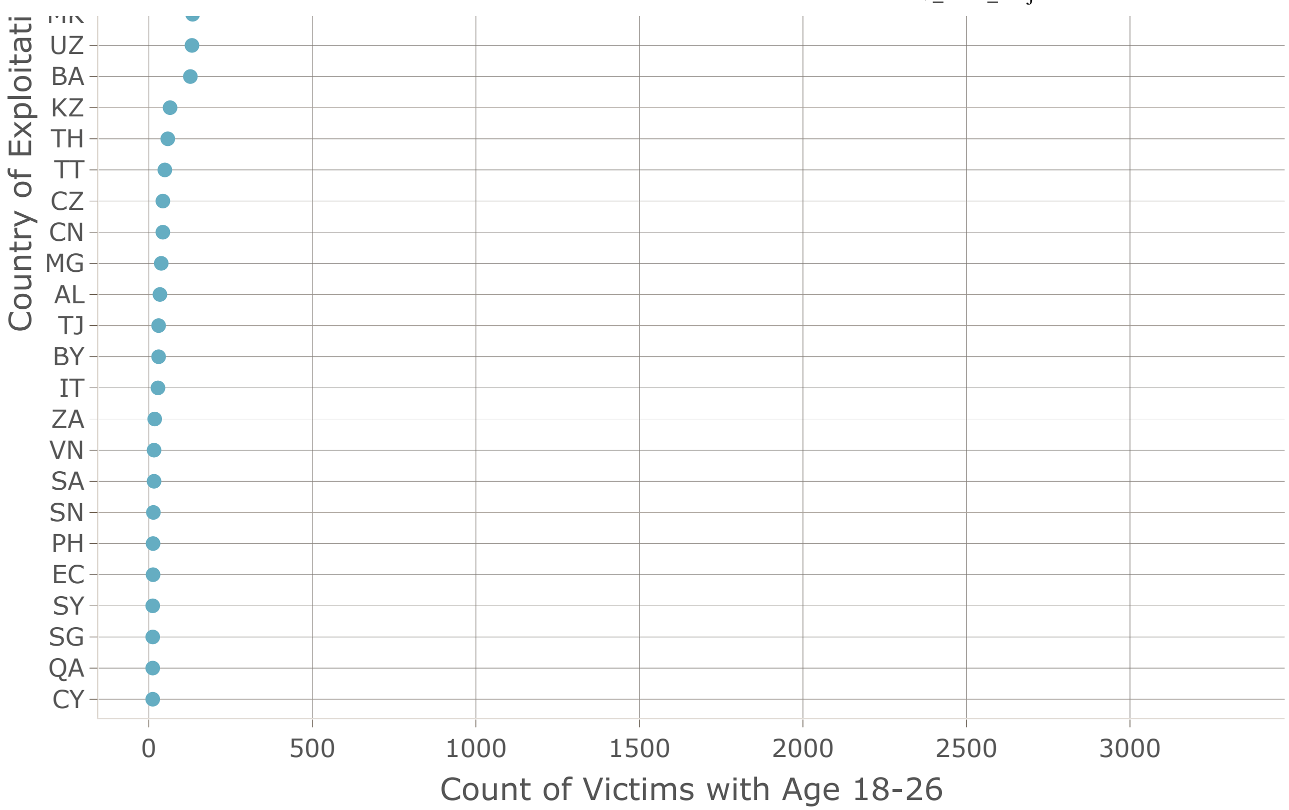
Recruiter Relation by Industry





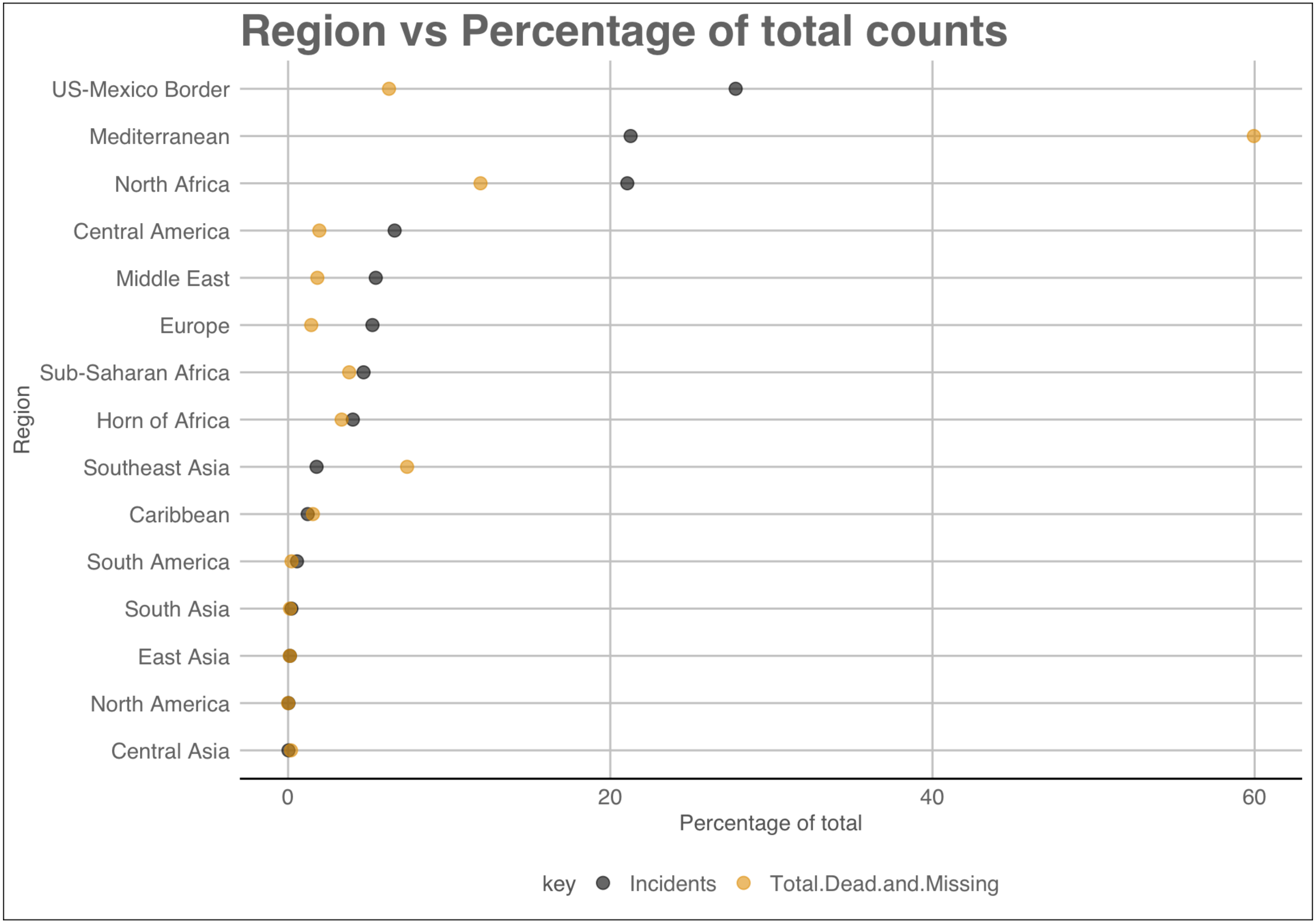
The grpah shows the relationship of the recruiter/trafficker(i.e the person who coerces/trafficks the victims) with the victim in different sectors or industries. It can be seen that a large percent(approximately 25%) of the vitims already know the perpetuator of the crime. The trafficker (for a major portion) is a family memeber or a friend. A significant amount of victims are trafficked by their intimate partner. This percent of known recruiters are greater in the hospitality sector than the other sectors. The second is that of domestic abuse.



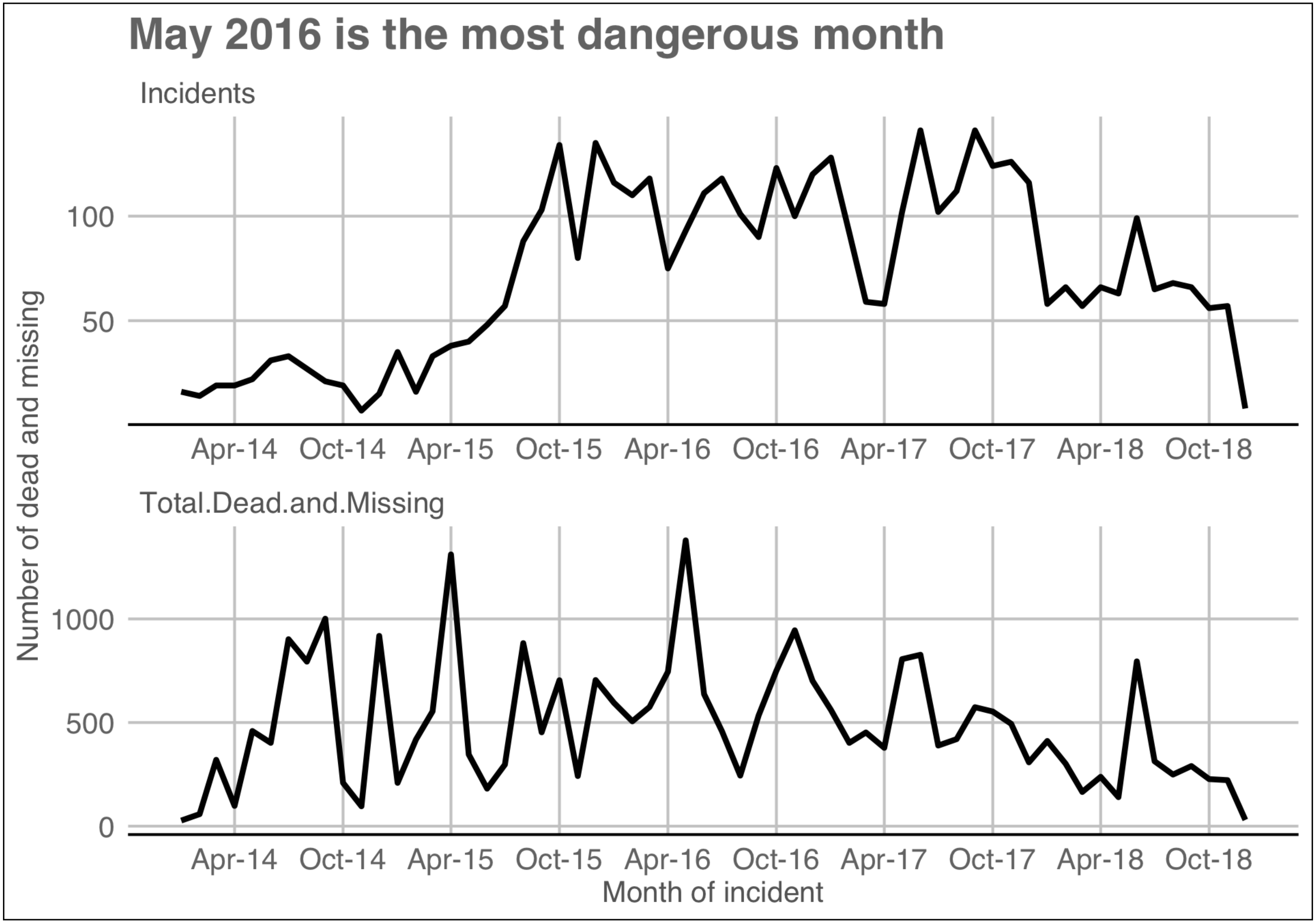


This graph shows the number of victims of the Age group 18-26 in different countries. This shows that the countries: United States, Ukraine, Maldova, Russia, Indonesia have the highest number of cases of exploitation. The count of victims in top three countries are a lot larger than the other country counts.

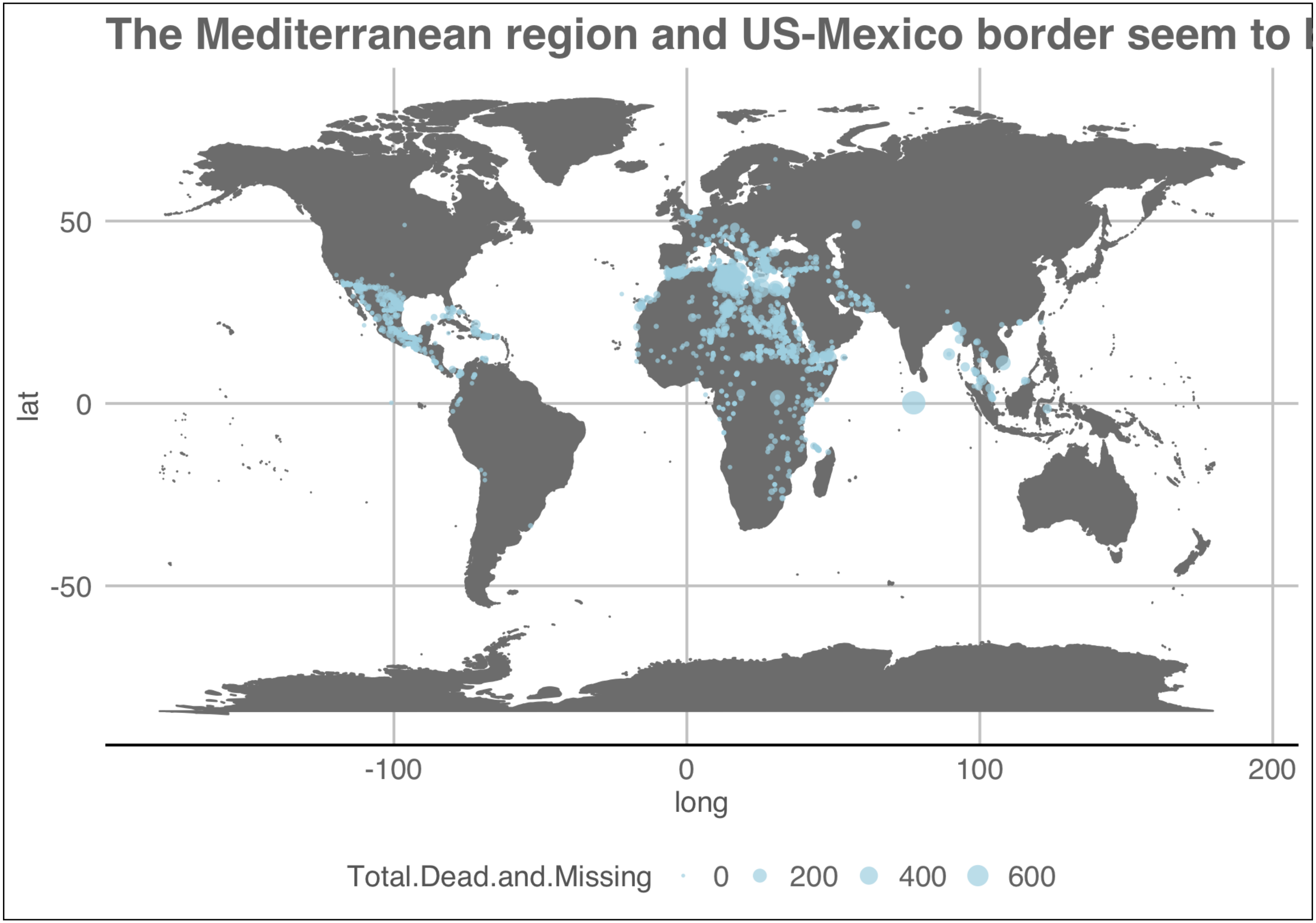
5.2 Missing Migrants dataset



The US-Mexico border has the most migrant incidents, but that border does not have the most fatalities. Instead, the Mediterranean region has the most fatalities. It is also not far behind in terms of the number of incidents. An even more surprising part is that Southeast Asia does not have that many migrant incidents, but a lot more fatalities with respect to the migrant incidents.



There are huge peaks in the number of fatalities in June 2016 and April 2015, but the number of incidents do not correspond to the same peaks, which may show that there were some incidents that took place that month which had a lot of people involved. Also, another promising fact is that as time goes on, we can see that the number of fatalities are decreasing over time, and so are the number of incidents.



The US-Mexico border and the Mediterranean area is the most active area in terms of migrant incidents. Most of the incidents correspond to the border of a lesser privilaged country to a greater privilaged country.

6. INTERACTIVE COMPONENT

The trafficking by country interactive component can be used to analyse trends and statistics in trafficking by country. Victims of trafficking are mapped by region. Each circle represents victims of trafficking in a country. The size of the circles represent a range of number of victims. Refer to legend to comprehend number of victims in each region. Red circles represent victims in the country of exploitation. Blue circles represent victims in country of citizenship. Countries may have a red and blue circle if they have victims coming into the country and going out of the country.

The trafficking by time component is an animation of trafficking by year. Each blue circle represents 10 victims of trafficking at country of citizenship. Each red circle represents 10 victims of trafficking at country of exploitation. The circles transition from blue to red as the victims move from country of citizenship to country of exploitation. Each year is represented by a 5 second interval. The number of victims and the countries involved in trafficking change over time.

Link for the Interactive Component : https://bl.ocks.org/akshatapatel/raw/97326ed42c9d5e29ca6986e87d656146/index.html (https://bl.ocks.org/akshatapatel/raw/97326ed42c9d5e29ca6986e87d656146/index.htm

Link to the gist : https://gist.github.com/akshatapatel/97326ed42c9d5e29ca6986e87d656146 (https://gist.github.com/akshatapatel/97326ed42c9d5e29ca6986e87d656146)

7. CONCLUSION

7.1 Future Directions

Some future directions might be to link both of the data sources together to get a list of possible illegal migrant incidents, or a possible variable which can distinguish in legal and illegal migrantion incidents.

7.2 Limitations

- The datasets are limited to certain regions and are have not been expanded to further regions of the world. Having data on more regions might allow us to look at more intra-migration incidents and more trafficking events.
- The data is very limited in time, ie the data does not have enough records for data which is old.

7.3 Lessons Learned

- Learning D3
- A major proportion of trafficked victims are young females.
- The most common form of exploitation of trafficked victims are sexual abuse(for females) and forced labour(for males).

- Most of the migration incidents take place in the Mediterranean and the US-Mexico border.

Link to the Github repository: https://github.com/ujjwal95/migration_and_human_trafficking (https://github.com/ujjwal95/migration_and_human_trafficking)