

Data Set analysis

Syracuse University iSchool

[Document subtitle]

Data Set analysis

**Group 4:**

**Robert Eason**

**Kristen Logue**

**Jack O’Connor**

**Jaci Willoughby**

Data Set analysis

Syracuse University iSchool

[Document subtitle]

**Syracuse University iSchool**

**IST 687 – Data Analytics**

**Winter 2019**

**Table of Contents**

[Data Set Description 2](#_Toc27581388)

[Project Scope 2](#_Toc27581389)

[Project Deliverables 2](#_Toc27581390)

[Data Cleaning 3](#_Toc27581391)

[Question #1 – What are the most dangerous routes? 4](#_Toc27581392)

[Question #2 – What is the most common way migrants die? 8](#_Toc27581393)

[Question #3 – Is it possible to predict deaths in the future? 10](#_Toc27581394)

[Question #4 – What can be done to prevent deaths in the future? 14](#_Toc27581395)

# Data Set Description

The Missing Migrant Dataset was found on Kaggle.com. It was derived from information gathered by The Missing Migrant Program beginning October of 2013. The dataset tracks the deaths of migrants and refugees who have gone missing along worldwide migration routes. The Missing Migrant Project has since become a recognized and valuable resource of information for Governments and Media outlets all over the world.

The data is gathered from a variety of sources and organizations, depending on the location of the migration route. Information has been shared by the Coast Guard, Medical Examiners, Media Sources and from interviews with surviving immigrants.

The International Organization for Migration (IMO) defines a migrant as any person who is moving or has moved across an international border or within a State away from his/her habitual place of residence, regardless of:

    (1) The person’s legal status;

    (2) Whether the movement is voluntary or involuntary;

    (3) What the causes for the movement are; or

    (4) What the length of the stay is.

The Data Set counts migrants who have died or gone missing at the external borders of states, or in the process of migration towards an international destination. It does not include deaths that occur in immigration detention facilities, during deportation, or after forced return to a migrant’s homeland.

# Project Scope

The scope of this project is to analyze the data on missing migrants to predict future deaths in an effort to prevent them from happening. The data shows us the migration routes and the types of deaths that are occurring. Using predictive measures, we could suggest to Border Patrol, the Coast Guard and other worldwide organizations what to look out for so they can find better ways to help the immigrants and refugees.

# Project Deliverables

Our goal is to answer four questions:

1. What are the most dangerous routes?
2. What are the most common ways migrants die?
3. Is it possible to predict deaths in the future?
4. What can be done to prevent deaths in the future?

To do this we:

* Cleaned the data to prepare the dataset for further analysis.
* Found and coded attributes of each region, route and death using linear regression.
* Created various charts to illustrate these results.
* Made predictions of future deaths that allowed us to develop actionable insights.

# Data Cleaning

In order to clean our dataset, we first had to read the data into an R Markdown from an excel sheet. Next, we sorted the data by vector name. All other NA’s in columns were replaced by zeros. These columns include: Total Dead, Number of Females Dead, Number of Males Dead, Number of Children Dead and Number of Survivors. We chose to replace these NA’s with 0’s because if those columns do not have a number in them it is because there were none of this type of person found. Finally, we removed columns on Information Source and Rating because they were not relevant to our data analysis.

A snapshot of our data:

A screen shot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Our cleaning code is:

library(readxl)

# READ IN DATA

miss\_mig <- read\_excel("Syracuse/IST687/Project/Missing\_Migrants\_Dataset\_2019OCT

0.xlsx")

View(miss\_mig)

# SORT BY VECTOR NAME [Z] THEN [X]

miss\_mig[with(miss\_mig, order(miss\_mig$`Reported Year`, miss\_mig$`Reported Month`, miss\_mig$`Region of Incident`)),]

# USE MEAN TO REPLACE NA GROUPING BY REGION OF INCIDENT

library(tidyverse)

new\_df <- miss\_mig %>% group\_by('Region of Incident') %>%

mutate\_all(funs(ifelse(is.na(.), mean(., na.rm = TRUE),.)))

# REPLACE ALL OTHER NA'S WITH 0 (ZEROES)

miss\_mig[is.na(miss\_mig)] <- 0

# REMOVE UNUSED COLUMNS

miss\_mig\_cln <- miss\_mig[-c(14,15,18,20)]

miss\_mig\_cln

# Question #1 – What are the most dangerous routes?

The data shows us that the most deaths by route are through the Mediterranean, North Africa and Sub-Saharan Africa. To get these results we first installed the SQL package, then determined the number of men, women and children who died in each region. We also wanted to look at the total number of dead or missing, including the UNKNOWN gender and age. We then looked at the percentage of death and missing per route compared to the total number of migrants and got a totally different story. This information shows us that North-America, US-Mexico Border and Central Asia Routes are the most dangerous, based on the percentage of reported migrants who died on this route. Finally, we created a World map to show the location of deaths reported. This map is a strong visualization tool to show where migrant deaths are being reported.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Migration Route** | **Women** | **Men** | **Children** | **Total # of Dead w/ Unknown** |
| Mediterranean | 811 | 962 | 408 | 18,995 |
| North Africa | 265 | 122 | 106 | 4,461 |
| Sub-Saharan Africa | 207 | 63 | 71 | 2,241 |
| Southeast Asia | 150 | 83 | 89 | 2,227 |
| US-Mexico Border | 97 | 22 | 13 | 1,832 |
| Horn of Africa | 70 | 113 | 37 | 1,182 |
| Middle East | 60 | 23 | 40 | 666 |
| Central America | 39 | 48 | 17 | 621 |
| Caribbean | 38 | 42 | 5 | 491 |
| Europe | 35 | 68 | 17 | 490 |
| South Asia | 18 | 6 | 5 | 289 |
| South America | 13 | 8 | 2 | 99 |
| East Asia | 1 | 2 | 0 | 52 |
| Central Asia | 0 | 0 | 0 | 31 |
| North America | 1 | 0 | 0 | 2 |

library(sqldf)

sqldf('Select "Region Of Incident",

sum("Number of Females") AS "Women",

sum("Number of Males") AS "Men",

sum("Number of Children") AS "Children"

from MMDS

WHERE "Number of Females" > 0

GROUP by "Region of Incident"

ORDER by sum("Number of Females") DESC

')

sqldf("SELECT

region\_of\_incident,

sum(total\_dead\_and\_missing) AS NUMBER\_OF\_DEAD\_MISSING

FROM dfmiss\_mig

WHERE total\_dead\_and\_missing <>''

GROUP BY region\_of\_incident

order by sum(total\_Dead\_and\_missing) DESC

")

|  |  |
| --- | --- |
| **Migration Route** | **% Dead & Missing** |
| Mediterranean | 31% |
| North Africa | 82% |
| US-Mexico Border | 89% |
| Southeast Asia | 73% |
| Sub-Saharan Africa | 64% |
| Horn of Africa | 37% |
| Central America | 32% |
| Caribbean | 62% |
| Middle East | 45% |
| Europe | 57% |
| South Asia | 83% |
| South America | 61% |
| Central Asia | 91% |
| East Asia | 56% |
| North America | 100% |

library (formattable)

MMDS1 <- sqldf('Select "Region Of Incident",

sum("Total Dead and Missing") AS "Dead and Missing",

sum("Number of Survivors") AS "Survivors"

from MMDS

WHERE "Total Dead and Missing" > 0

GROUP by "Region of Incident"

ORDER by sum("Total Dead and Missing") DESC

')

MMDS1

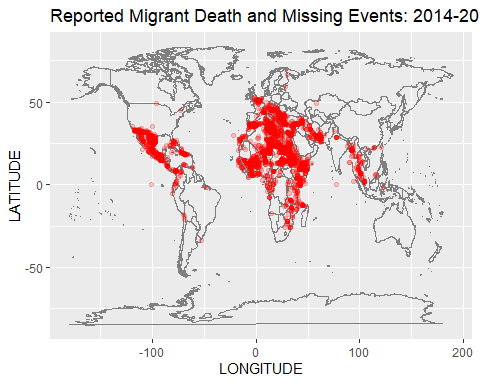
MMDS1$"Total Migrants" <- MMDS1$`Dead and Missing` + MMDS1$`Survivors`

MMDS1$"DMpercent" <- MMDS1$'Dead and Missing'/MMDS1$'Total Migrants'

MMDS1$"DMpercent" <- percent(MMDS1$"DMpercent")

order(MMDS1$DMpercent)

MMDS1

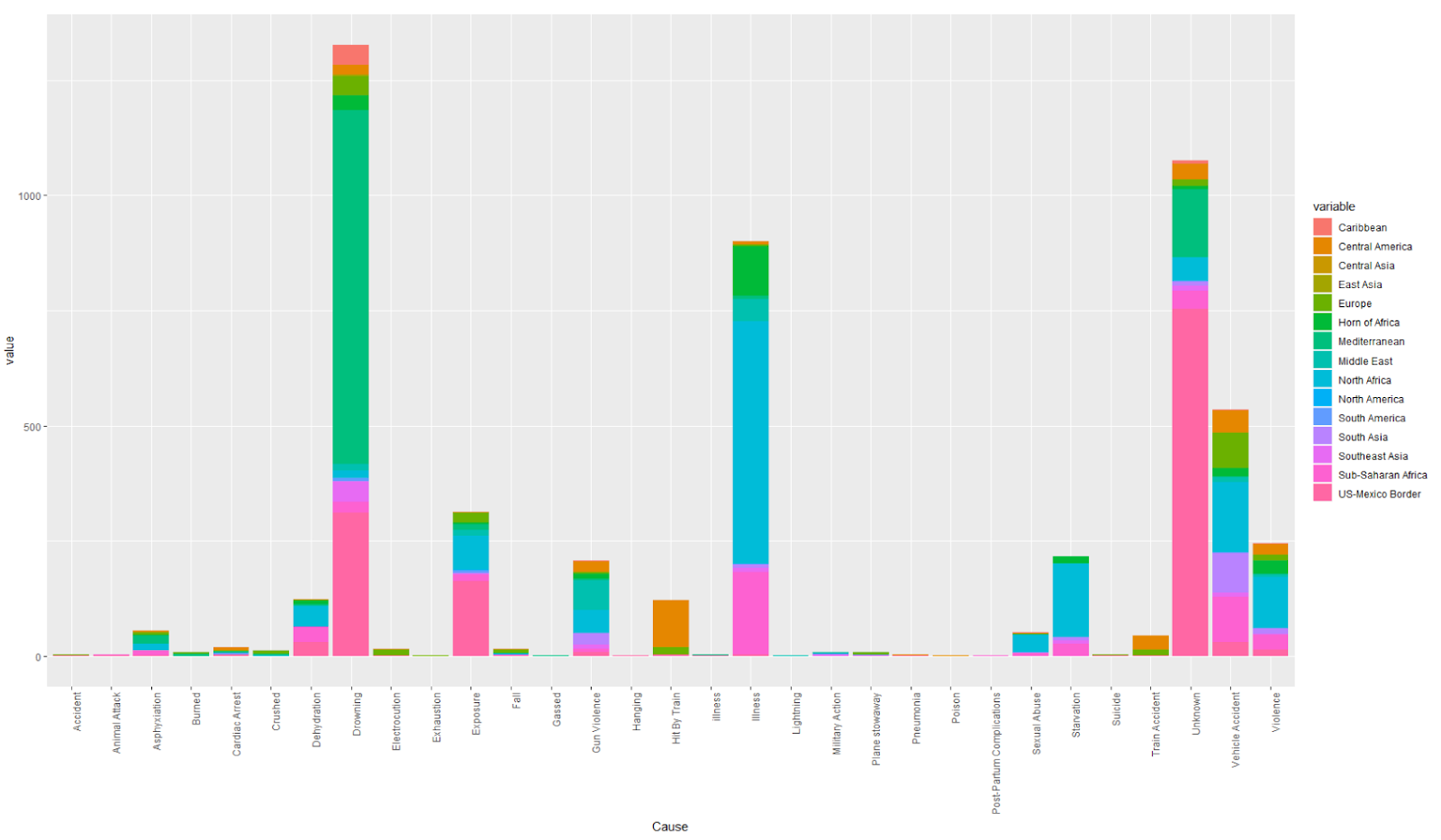


# Question #2 – What is the most common way migrants die?

Next, we wanted to look at circumstances surrounding each death. The data needed to be cleaned further for this to create “categories” for cause of death. A sub-data set was saved so as not to hinder the original data, and cause of death was substituted with the category of death that best fit – as the data came from a variety of sources and no standards of language were used. This allowed for easier analysis on the type of death, as well as the creation of a visualization.

The results indicate that Drowning is the #1 cause of death of migrants in the study, followed by unknown causes, then illness that went untreated. Vehicle Accidents were higher than Exposure. Violence and Gun Violence were almost equal, followed by starvation, train accidents, and sexual assault.

Using our Cause table, we created a stacked chart to illustrate our results:



Our data cleaning code for this question:

The columns not needed for this Sub-Dataset were removed, and cause of death substitutions were made:

CauseSub <- MMDS[ -c(1, 3:12, 14:16)]

Here is a sample of the code used to change the Cause of Death to a Category:

CauseSub$`Cause of Death`[CauseSub$`Cause of Death`== "Accident (non-vehicle)"]<-"Accident"

CauseSub$`Cause of Death`[CauseSub$`Cause of Death`== "Plane stowaway"]<-"Plane stowaway"

CauseSub$`Cause of Death`[CauseSub$`Cause of Death`== "Killed by crocodile"]<-"Animal Attack"

CauseSub$`Cause of Death`[CauseSub$`Cause of Death`== "Killed by hippopotamus"]<-"Animal Attack"

CauseSub$`Cause of Death`[CauseSub$`Cause of Death`== "Killed by hippoptamus"]<-"Animal Attack"

CauseSub$`Cause of Death`[CauseSub$`Cause of Death`== "Envenomation"]<-"Animal Attack"

We needed to reshape the data once it was changed, so we installed the reshape2 package which allowed us to easily change the data into a data frame:

library(reshape2)

Cause <- dcast(CauseSub, CauseSub$`Cause of Death` ~ CauseSub$`Region of Incident`)

names(Cause)[1] <-"Cause"

Cause

# Question #3 – Is it possible to predict deaths in the future?

We decided that the different “number of dead” variables were strong data points that we could attempt to predict. By being able to predict these variables we can explore some possible prevention options to keep migrants safe in the future. In order to predict future deaths, we first had to clean our data. We sub set the data to include only the information we needed, built a linear model and pulled a summary of that model. The summary showed us that, although our P value is low, the variance in our outputs cannot be explained by our inputs. Next, we created a residual plot, tried to train an SVM model and created a neural network visualization.

The data of Number of Dead is largely centered around values between 1 and 10, with some extreme outliers in the hundreds. This makes it very difficult to fit a model to the data to predict future deaths.

Here is a deeper look at the process we used to answer this question:

First, we cleaned the data. We started by sub setting the data to just the variables we needed for building our models:

svm\_data <- subset(missing\_mig, select = c(“Region.of.Incident”, “Reported.Date”, “Number.Dead”, “Minimum.Estimated.Number.of.Missing”, “Number.of.Survivors”, “Number.of.Females”, “Number.of.Males”, “Number.of.Children”,

“Cause.of.Death”, “Migration.Route”))

Next, we built a linear model:

lin.mod <- lm(formula = Number.Dead ~ Number.of.Survivors” + Number.of.Females + Number.of.Males + Number.of.Children , data = svm\_data)

Summary(lin.mod)

A screenshot of a cell phone

Description automatically generated

This result is the summary of only our most statistically significant variables. All insignificant inputs were removed. Our P values are all low for each of the independent variables, which is probably because the numerical variables all describe the constitution of each group (i.e. number of men, women, and children). Even though the P values are small, the adjusted R^2 proves that our inputs do not explain our outputs. Only 6% of the variance in our output variable is explained by variance in the inputs.

We pulled a residual plot of our linear model to see what it looked like:

A screenshot of a social media post

Description automatically generated

Next, we trained an SVM model using the “kernlab” package. We divided our data into a training set and a testing set. Next, we trained our SVM model:

nrows <- nrow(svm\_data)

sample\_size <- round(nrows \* .70)

training\_index <- sample(1:nrows, size = sample\_size, replace = FALSE)

svm\_train <- svm\_data[training\_index, ]

svm\_test <- svm\_data[-training\_index, ]

library(kernlab)

ksvm\_model <- ksvm(Number.Dead ~ ., data = svm\_train, type = “eps\_svr”, kernel = “vanilladot”) #Training error = .823517

pred1 <- predict(ksvm\_model, newdata = svm\_test)

svm\_test$pred1 <- pred1

svm\_test$error <- svm\_test$Number.Dead – svm\_test$pred1

A close up of text on a black background

Description automatically generated

As with the linear model, our SVM model is not as good at predicting number of deaths as we had hoped. As pictured above, in some instances we get a close prediction, but in others we are off by more than 150 deaths.

Finally, we wanted to see if a neural network would offer more predictive power than our previous models. We used the “neuralnet” package to build and train this network. Our neural network was not any better at predicting number of deaths than our other models. Our error on the neural network is extremely high and offers little in prediction efficacy.

A picture containing text, map

Description automatically generated

library(neuralnet)

neural\_network <- neuralnet(Number.Dead ~ Number.of.Survivors” + Number.of.Females + Number.of.Males + Number.of.Children, data = svm\_train, hidden = 4, rep = 1, lifesign = “minimal”, linear.output = TRUE, threshold = 2)

# Question #4 – What can be done to prevent deaths in the future?

Although we cannot seem to predict future deaths based on the data we have from the Missing Migrants Project, we can conclude from both our route and cause of death data that there are migration routes that are more and less dangerous. Using this data, we can suggest alternative routes for migrants (if they are reachable) or tell patrol agents where they should be concentrated to better help the migrants who need it.

According to our data, the most deaths are on routes through the Mediterranean, North Africa and Sub-Saharan Africa. This tells us that these three migrant routes, due to the sheer volume of deaths, need more people and a larger amount of resources working to protect those migrating through. When you look at the death rate by route as a percentage of reported migrants you see that North-America, US-Mexico Border and Central Asia Routes have the highest percentage of people dying while traveling them. Although the number of people dying on these routes overall is smaller, the high percentage tells us that North-America, US-Mexico Border and Central Asia Routes are the most dangerous routes to travel and should be avoided by migrants. By spreading information on route alternatives, we could save migrant lives.

The most common migrant death is drowning. Based on this, we can conclude that routes that include travel over water are riskier than those that don’t. One way to prevent deaths of this type would be for the Coast Guard, Boarder Patrol and other applicable service members to have a more significant presence on water routes to help potential victims. The third most common cause of death is illness that went untreated, showing us that migrants need access to medical assistance throughout their route. Providing information on where safe, accessible medical assistance can be found on these routes could give migrants easier access to life saving care. The second most common cause of death is “unknown”, which tells us there is still a lot of research to be done on these migrant routes.

Spreading information on safer migration alternatives, providing quick rescue efforts and dispatching the correct medical assistance will be key in preventing deaths along any migrant route.