Homework 1: Exploratory Data Analysis on the El Salavordan Conflict

Due February 19th, 2021 at 5 PM EDT

General instructions for homeworks: Please follow the uploading file instructions according to the syllabus. Your code must be completely reproducible and must compile.

Advice: Start early on the homeworks and it is advised that you not wait until the day of as these homeworks are meant to be longer and treated as case studies.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml. No late homework's will be accepted.

Total points on assignment: 5 (reproducibility) + 10 points for the assignment.

El Salvador Civil War

Recall that between 1980 and 1991, the Republic of El Salvador witnessed a civil war between the central government, the left-wing guerrilla Farabundo Marti National Liberation Front (FMLN), and right-wing paramilitary death squads. After the peace agreement in 1992, the United Nations created a Commission on the Truth (UNTC) for El Salvador, which invited members of Salvadoran society to report war-related human rights violations, which mainly focused on killings and disappearances.In order to collect such information the UNTC invited individuals through newspapers, radio, and television advertisements to come forward and testify. The UNTC opened offices through El Salvador where witnesses could provide their testimonials, and this resulted in a list of potential victims with names, date of death, and reported location.

In this assignment, you will explore the UNTC data set to get a better understanding of how to work with real data versus toy data. Let's read in the data.

```
library(knitr)
library(RecordLinkage)
```

```
# read in data
df <- read.csv("../sv-mauricio/sv-mauricio.csv")
head(df)
```

```
##     X  ID         lastname firstname day month year geocode HandID dept muni
## 1  26  32 ASENSIO ERNANDES   ALBERTO  NA     2 1981  150000     NA   15   NA
## 2  84  95   PALASIOS AYALA    OBIDIO  NA    10 1985  150000     NA   15   NA
## 3 100 117            PALMA SEBASTIAN  13     5 1980   40000     NA    4   NA
## 4 143 173            PERES   ARCADIO  NA     8 1984   40000     NA    4   NA
## 5 170 205     MAYA QUESADA   ANTONIO  22     9 1984       0     NA    0   NA
## 6 189 227            MEJIA   ALFONSO  13     5 1980   40000     NA    4   NA
```

```
dim(df)
```

```
## [1] 5395   11
```

Task 0: Exploratory Data Analysis and Collecting Background Information

Perform an **exploratory data analysis** to help you understand this case study.

Write this up in a **short report** (maximum four pages), summarizing only the most interesting and important findings. Use tables, visualizations, and make sure that you code is **well documented** and **reproducible** so that anyone in the class could benefit from your analysis!

Some questions that you might want to ponder are the following:

- Do you need to do some background reading on El Salvador to understand the context of what your studying here? It might help to look up the munipalities and departments (and how these are defined in El Salavdor). How many are there? You might benefit from looking at a map of El Salvador as well. Can you map the munipalities to a map of El Salvador.
- What do the attributes (features) stand for?
- Do you have missing data? How much missing data do you have?
- Do you have unique identifiers?

Comment on how this assessment will help you be a better data scientist when working with this case study in the future. (Just a few sentences will suffice).

Task 1: Exact Matching

Apply exact matching, off-by-one matching, off-by-k matching to this data set and report the pairwise precision and recall. What do you find?

Task 2: String distances

What types of string distance metrics are appropriate and which are not appropriate?

Task 3: Decision rules

How would you build a decision rule for matches/non-matches based upon scoring rules. What would your scoring rule be? Write this up.

Task 4: Implementation/evaluation

Implement your proposal for scoring based rules and test this our providing the pairwise precision and recall.

Task 5: Insights and reflections

Give insights into how you might be able to improve deterministic approaches moving forward if you re-did your analysis.