PROJECT TITLE: Regis University Practicum I – Capstone Project: A Clustering and Analysis of Terrorist Attacks

PROJECT SUMMARY

I am wanting to address one of two scenarios: I am wanting to perform a clustering example on the Global Terrorism Dataset to find relationships. Once I have a model that creates clusters, I will then do different tasks; I will perform an analysis on the clusters to see trends, relationships, and see if there are any defining characteristics. After I do this, I will then pass new, never before “seen” data points to the model to show which cluster they belong with.

MILESTONES (Bullet point STEPs in your project. You can tag “DONE” for things are done)

(Example: You can edit for yours. You can also make it more detailed if you want.)

* Data collection (DONE)
* Data cleaning (In Progress)
* EDA (Done) Might explore other insights as they come
* Mapping of data points (Done)
* Organization of final project (In Progress)
* Final write-up (In Progress)

This week’s progress:

I started and got a good chunk of the exploratory analysis done. To me, EDA is an ongoing process, so I have marked it done but I will continue to look for interesting trends and patterns to help make sense of the clusters that will eventually be created.

Issues and Discussions:

I did run into one issue. One of the weapon categories is “Vehicle (non to include vehicle-borne explosives, i.e. car or truck bombs)” which is an incredibly long title especially when I am trying to put it into a visualization. I tried looking for something along the lines of (in R) gsub(“Long title”, “Short Title” Weapons column) but ran into that there were spaces and special characters (like parentheses) and it would not substitute correctly. So for the time being I have moved on, but I would love some help on how to substitute a specific data value (remember it is not a column, but a specific value within a column, google had a hard time differentiating this). I did see that it might be a regex solution?

To Do

* Feature Selection
* Classification using K-nearest neighbors
* Classification using hierarchy
* Organization on Github
* Data Library
* Compare and Contrast classification models

Resources

Mapping - https://ourcodingclub.github.io/2016/12/11/maps\_tutorial.html

Week 2 Progress Report

MILESTONES (Bullet point STEPs in your project. You can tag “DONE” for things are done)

(Example: You can edit for yours. You can also make it more detailed if you want.)

* Data imputation is in progress.
* I have used two different methods to impute missing data, one was using the mice package and was using a random forest. I could not get the random forest one to work because the work load was too much and no progress was being made. With the MICE package, I had to run it with only 5 iterations and with it only creating 3 datasets.
* Because the imputation is taking so long, hours at time, and I had to restart my R session a couple of time, I have not been able to actually do my PCA.
* I did some research and so I believe that I will be able to do my PCA either tomorrow or early next week.

This week’s progress:

I got my data down to where I am just looking at numeric data for the time being. My data includes categorical data and numbers attached to it. So one column is numbers 1-9 and each one of those corresponds with a text variable in the column next door. I also only kept variables that had over 80% of the values in them. Between keeping only numeric columns and ones with 80% or more of actual values, my dataset is now around 41 columns. This is still making the imputation process quite long, but I do think that I will be able to do PCA quite easily once I get the NA’s taken care of and I have complete cases to work with.

Issues and Discussions:

With my PCA and imputing, I kept just the numeric values, but now I am wondering if having numbers 1-9 mean something and not just actual numbers. That one categorical variable (say White/Caucasian is the number 1 and Hispanic is the variable 2) it does not make sense that 2 is two times greater than 1, or that Hispanic is twice as much as White/Caucasian. So I will do some research on this.

To Do

* Feature Selection
* Research how to handle categorical numeric values
* Classification using K-nearest neighbors
* Classification using hierarchy
* Organization on Github
* Data Library
* Compare and Contrast classification models

Resources

Mapping - <https://ourcodingclub.github.io/2016/12/11/maps_tutorial.html>

Principal Component Analysis - <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/>

Imputation Strategies - <https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17>