1. Name, Contact info (e.g. email/phone).

Ryan M. Allen

719-239-0607

[Rmallen09@yahoo.com](mailto:Rmallen09@yahoo.com)

[Linkedin](https://www.linkedin.com/in/ryan-allen-b66a8069/)

[Twitter](https://twitter.com/AllenRyan91)

[Github](https://github.com/RyanMAllen)

1. Title of the project

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

1. High level description of the project: what question or problem are you addressing?

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.

1. What type of data science task is it? (some example answers but not limited to)
   * Clustering using unsupervised learning
   * Principal Component Analysis to assist in feature selection
   * Exploratory Data Analysis
   * data visualization
2. Data: Brief description of data. How big do you expect the data will be? Is amount of your data too big or too small? If you're web-scraping or collecting data, how long do you expect to collect the data?

The dataset comes from [Kaggle](https://www.kaggle.com/START-UMD/gtd) and it includes 135 columns and has been collected since 1970-2017 with the exception of 1993. I think it will be a manageable size of data, with that many features though, there will be quite a bit of feature selection techniques that will need to be used to get the 100+ columns into a manageable number.

The data contains information on the perpetrators of the attack, targets, outcome, type of attack, location details, country where it occurred, region, target nationality, name of the group that attacked, casualties, number of terrorists involved, weapon type, etc… The dataset has a wide variety of datatypes, from geographical data, to text (descriptions of the attack, group, factions, etc…) and numbers.

1. How will you analyze the data? What machine learning methods do you plan to use, and/or what business intelligence aspect do you plan on incorporating?

Before I get into the clustering, I will perform PCA to limit the number of features that I am handling and using in my model. After that, I am planning on using hierarchical clustering and/or K-means to see if there are any relationships/connections/groupings that emerge from the terrorism dataset. This could be used to see if there are unique characteristics amongst the data that are not easily seen. Once I cluster the data I will then run some analysis on the strength of the clustering. I have read that there are several ways to see the similarities in clustering and I will employ these techniques to see the relative separation of the clusters.

Once I have the clusters I will then assign each data point a cluster and perform a classification on new data points to see which cluster they should belong to.

1. Describe any anticipated difficulties and problems. Discuss how you may overcome the problems.

Anticipated problems

* + - I think the feature selection could be challenging. I know that there will be some features that I will most likely have to manually exclude without much evidence or strong reasoning behind it, but I will research the best ways to limit features when you are dealing with this many columns.
    - One other problem that I am anticipating running into is how to handle the text columns. For most of the columns there is a numeric value or single letter value and a corresponding column with text (so 1 means United States, 2 means Canada, etc…). What I might have to do is keep only the numeric columns for the model and once I have the model created, for the analysis, visualizations, and write-up substitute in the text columns to make more sense of the data. It will require a good amount of organization and attention to the details.
    - I am not worried at the moment that I have bit off more than I can chew but I will see if I have time to compare and contrast the K-means with the Hierarchical Clustering AND classify data that I have withheld from the model.

1. Suggest a timeline for the project.  This should be a weekly breakdown of what you plan on doing each week.

Week 1 – Proposal, research clustering

Week 2 – Exploratory Data Analysis, research best ways to compare and contract clustering models

Week 3 – PCA

Week 4 – Create the hierarchical clustering example and analyze/visualize the clusters

Week 5 – Create the k-means clustering example and analyze/visualize the clusters

Week 6 – Compare and Contrast the models, choose one, assign clusters to the original dataset and run classification problem

Week 7 – Write up and general organization of the project

Week 8 – Finish write up, presentation, and submit.

Some of the weeks have less “work” then others and can be combined to accommodate for other weeks needing more time and “work”.