## ITEC657 – Project Proposal

**Group A (Tue 7pm)**

Jake Patrick Ebreo (45675406), Suhash Reddy Immareddy (45693242)

Pornpawee KHONCHOHO (45480397), David Warren (45939225)

## **Predicting Crime Rates Using Temporal Patterns and Environmental Conditions**

This project will analyze patterns of reported criminal behavior in the city of Washington, D.C. First, we will provide a descriptive analysis of the types and relative rates of crime in different areas to identify potential clusters in the data. Then we will provide predictive models to estimate the number of crimes expected on a given day based on the temporal factors (day, date and season), environmental factors (temperature, precipitation, snowfall) and economic indicators (current unemployment rate and trends).

**Problem:**

Rates of criminal behavior are influenced by a huge number of social, environmental and economic factors and can fluctuate significantly from day-to-day. This unpredictability means it can be difficult for police departments and public services that deal with criminal behavior to be appropriately resourced where and when they are required.

.

**Goals:**

1. Provide visualizations and analysis of the crime occurring in different areas of the city and create area profiles using cluster analysis to allow law enforcement to identify commonalities between areas. These commonalities can enable law enforcement to make predictions about trends in these areas and inform decisions regarding which strategies may be viable in which areas. A strategy that is effective in a particular area is most likely to be applicable to other areas in the same cluster or profile.
2. Create a reliable predictive model to assist in proper and efficient allocation of law enforcement resources. A model that can give estimates of the total volume of crime to expect allows enforcement agencies to improve police response time, engage in pro-active prevention campaigns and better manage the stress and fatigue of officers.

**Datasets:**

**Felony Crime Incidents – Washington D.C**

* Source – City of Washington, D.C.
* Link – [https://opendata.dc.gov](https://opendata.dc.gov/)
* Format – CSV
* Features of interest – Offense, Date, Precinct, Location (longitude, latitude)
* Work needed:
  + Merge Data Sets - multiple year reports into a single dataset
  + Remove incomplete and NA entries.
  + Group by date, time and area to be used in different aspects of the analysis.

**Environmental Data – Washington Arboretum Weather Station 2010 - 2018**

* Source - National Center for Environment Information
* Link - <https://www.ncdc.noaa.gov/cdo-web/>
* Format – CSV
* Features of interest - Temperature Min and Max, Precipitation, Snowfall, Wind Speed, Date.
* Work needed:
  + Merge Data Sets - multiple year reports into a single dataset
  + Calculate a relative temperature measure by comparing daily values to monthly averages.

**Unemployment Rates**

* Source – U.S. Federal Government - Data.Gov
* Link – <https://data.gov>
* Format – CSV
* Features of interest: Unemployment Rate, Area (Washington D.C), Month.
* Work needed:
  + Refine data to relevant date range.
  + Join monthly unemployment figures to a daily date index so that they can be combined with other datasets.

**Public Holidays – Washington D.C. 2010 - 2018**

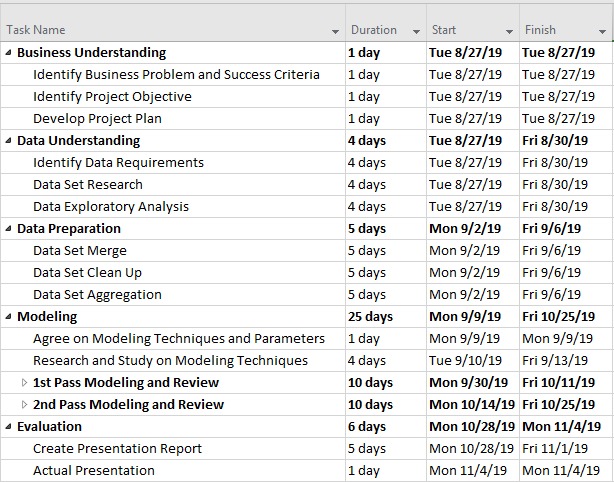
* Source - Calendarpedia
* Link - [www.calendarpedia.com](http://www.calendarpedia.com)
* Format - CSV
* Work needed:
  + Merge all holidays from 2010 to 2018.
  + Boolean encoding of public holidays to calendar days for analysis.

**Techniques:**

Heatmaps, clustered bar charts and choropleths will be used to perform initial visual analysis of crime in the city. Then K-Means clustering will be applied to areas to provide crime profiles. Cluster centers will be interpreted, and clusters will be visualized geospatially over the city.

A number of regression techniques will be applied to total crime and specific types of crime per day to produce predictive models including linear regression, random forest regression and gradient boosting regression. Features will be assessed for predictive value and model accuracy will be compared for different techniques, then the best model using the most important features will be recommended.

**Project Timeline & Milestones:**

****