# Capstone Project Proposal-1

# Subject: Data Science Career Track

# Student: Franklin Cid

**Problem:** (911 Police Calls for Service (Baltimore, MD) Data EDA)

# The 911 Public Safety Answering Point (PSAP) service area boundaries in California, according to the National Emergency Number Association (NENA), is a facility equipped and staffed to receive 9-1-1 calls. The service area is the geographic area within which a 911 call placed using a landline is answered at the associated PSAP. New functionalities on other systems developers now include calls from mobile phone and texting. This task will only address calls from landlines.

# Police-Calls-for-Service are either emergency or non-emergency. Based from the target dataset, there are a few columns that are categorical which answer several problems.

# Prospective Clients/Usage?

Prospective clients include systems developers (e.g. [West/EC@TS](mailto:West/EC@TS)), PSAPS (the facilities, usually on the county level, that actually receive and answer the emergency and non-emergency calls all over the U.S.), and local/city government.

This task is an exploratory data analysis project. As such, I want to find the answers to the problems below:

1. What is the percentage of emergencies and non-emergencies are from a dataset?
2. What is the percentage are High Priority, Medium, and Low are from the emergency calls?
3. What is the district with the greatest number of High Priority calls?
4. What is the type of emergency that's mostly incurred?
5. Is there a correlation of the call priority to a district?
6. Is there a correlation of the call type (description column) to a district?
7. Is there a correlation of the call type (description column) to a call priority?

# Data/Acquisition:

The data that will be used for this project comes from the City of Baltimore in Maryland. The data can be downloaded from the city website: data.BaltimoreCity.gov.   
(Rows = 4.03M; Columns=8; Each row is a Call)

With the answers that I'll find from the problems above, my clients/prospects will be able to:

1. Properly schedule their help desk personnel for peak calls, districts, etc.
2. Properly anticipate what kind of emergencies to anticipate – cops, fire, ambulance, etc.
3. Properly anticipate what call priorities are in a district.
4. Properly anticipate what types of calls are in a district.
5. Properly anticipate what types of calls are for a call priority.

# Solution Outline/Approach:

1. Download dataset from the data source.
2. Data wrangling
   1. Check and fix NULL values
   2. Check and fix 0 values
   3. Check and fix/remove error values
3. Graphs: With matplotlib, numpy, and Seaborn packages, use appropriate stat/plot functions to answer the following problems. {sum(), sns.swarm}
   1. Percentage of emergencies(EC)/non-emergencies(NC) – daily, weekly, monthly?
   2. Percentage of High Priority, Medium, and Low are from the EC – daily, weekly, monthly?
   3. District with the most High Priority call – daily, weekly, monthly?
   4. Type of emergency that's mostly incurred – daily, weekly, monthly?
   5. Correlation of the call priority to a district?
   6. Correlation of the call type (description column) to a district?
   7. Correlation of the call type (description column) to a call priority?
   8. Rank by districts
      1. heat map –use seaborn= sns.heatmap (2 var), pie chart, bar,

Features = e.g.: age, wt, ht, - compute correlation bet 2 var; represent all in a heat map

correlation bet ed/income, Happiness/wealth, #flights/timeOfYr

Ref: see blogs – how it’s used

1. Conclusion/Findings
2. **Recommendations:**

# I'd be interested to continue to work on this dataset to predict what type of calls, how many emergency calls, etc, in the future.

# Deliverables:

1. Python Code
2. Short Paper
3. Slide Deck

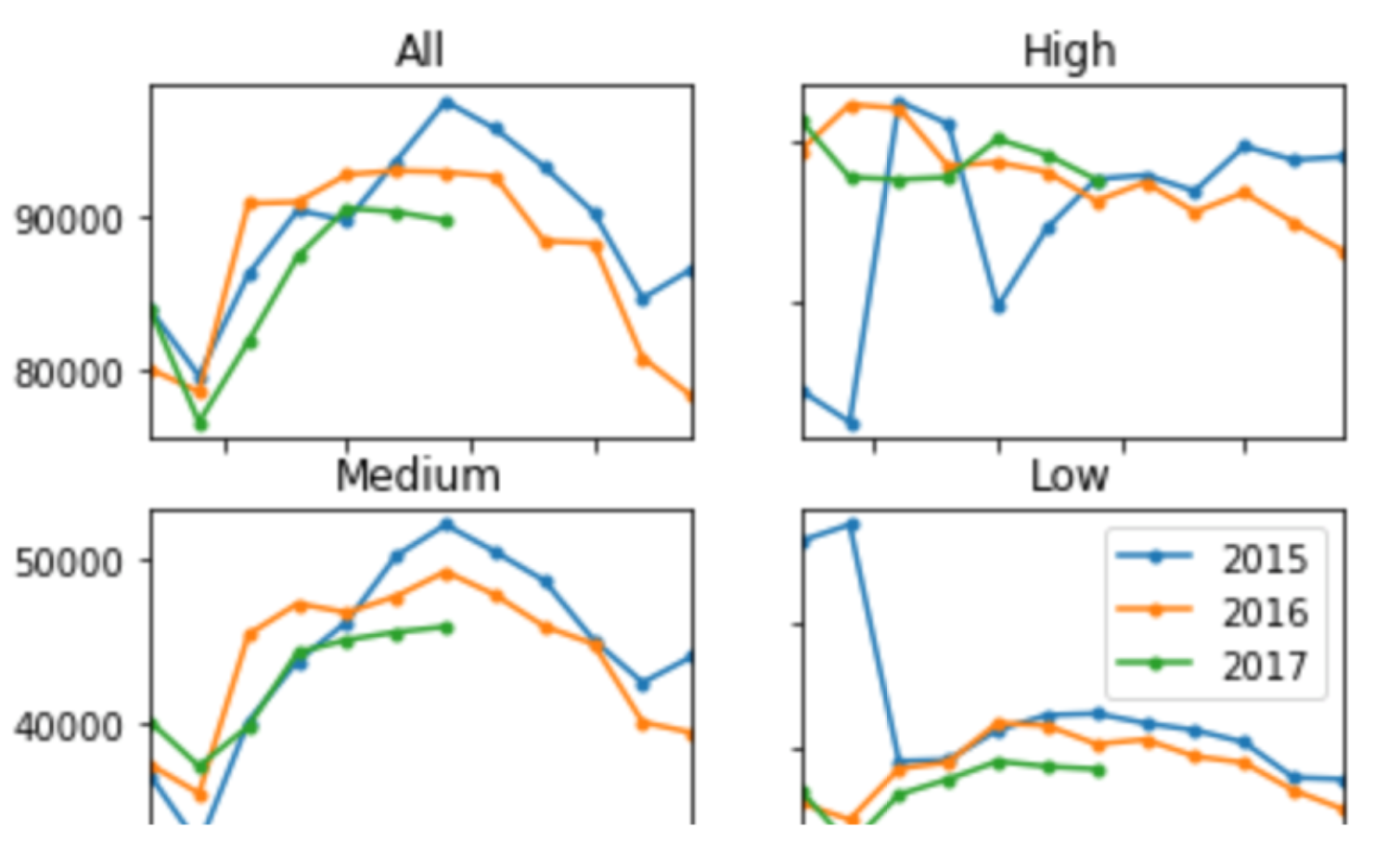
Notes:

3 models:

1. prediction – from history, predict future = tell system what it is/not and hope in the future it can recognize; continuous vars, house/stock/sales - supervised learning
2. classification – a pix of dog/cat, spam/not Spam - supervised learning
3. clustering – is news headline related to economy or sport or education? - unsupervised learning

Same dataset, train (train=80/test=20) a model based on some part of the dataset  
(generally, the more data, the better)

Sample Graph/Chart:



# Appendix:

PSAP = Public-Safety Answering Point

NENA = National Emergency Number Association (NENA)

# Data Source: <https://data.baltimorecity.gov/Public-Safety/911-Police-Calls-for-Service/xviu-ezkt>

# Data Columns: 3.86 mil rows (Milad: 10k, or 1-mo; cold=Jan/Feb and hot=Jul/Aug), 8 cols, 1 call/row robbery on bike, laptops

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Description | Type | Comment |
| recordId | System generated? | number | 2749202 |
| callDateTime |  | datetime | '08/10/2017 03:25:00 PM'  format: mm/dd/yyyy hh:nn:ss AM/PM |
| priority | Priority of emergencies | text | [Low, Medium, High] |
| district | District codes | text | [ND, ] |
| **description** | District Description | text | LARCENY |
| callNumber | “how is generated? Where’s the source? | text | P172221666 |
| incidentLocation |  | text | 3100 ST PAUL ST |
| Location | IncidentLocation + lat/long | Location (lat/long) | "3100 ST PAUL ST BALTIMORE, MD (39.326142, -76.615959)" |

Sample Data:

recordId,callDateTime,priority,district,description,callNumber,incidentLocation,location  
2749202,08/10/2017 03:25:00 PM,Low,ND,LARCENY,P172221666,3100 ST PAUL ST,"3100 ST PAUL ST  
BALTIMORE, MD (39.326142, -76.615959)"  
2785533,08/23/2017 12:53:00 PM,Medium,NE,911/NO VOICE,P172351302,3300 BELAIR RD,"3300 BELAIR RD  
BALTIMORE, MD (39.321622, -76.573597)"

