**Data Science Project Protocol**

*Author(s): Shiri Margalith*

# Introduction

Most Americans report they have been victims of some type of criminal act, at least once. Specifically, in the American cities, crime (of all types) is much higher than it is in other modernized societies. As a result, many Americans fear living in urban areas.

Specifically, in the Los Angeles metropolitan area, which is the second-largest metropolitan area in the US after New York, with over 10 million residents, the crime situation has gotten worse in the recent years. In 2016, for example, crime rate in Los Angeles rose by 5%, while crime across all of California declined by approximately the same amount. In this project we will review crime data ranging from 2010 through 2019 and will aim to understand the changes in crime rate in each district.

Building a safe community is one of the top priority of LA’s mayor and he has taken several steps to extend the LAPD Community Safety Partnership plan, whose purpose is to create a closer bond between the residents of Los Angeles and the police officers who serve them.

Since the plan is not restricted to pure police work and relates also to developing various activities that fit the community and will hopefully reduce crime, our project’s support to the plan is twofold:

1. Most crimes follow a pattern, especially when conducted by well organized gangs, as it is in Los Angeles in many cases. Understanding these crime patterns and criminal behavior (location, hour, type) will help recommend the best police force allocation in the Los Angeles area in order to prevent crimes / react quickly when those occur. In order to do so we will analyze the nature of the crimes, where and when they are most likely to take place and what are the characteristics of the criminals.
2. Being one of the most ethnically and culturally diverse counties in the US, it will be interesting to also look at the demographic characteristics of each Los Angeles district in order to better understand the population. This, in the future can be used to develop a taylor-made activity plan for each district.

The datasets that will be used for this project are the following:

* Arrest Data from 2010 to 2019 - [LINK](https://data.lacity.org/A-Safe-City/Arrest-Data-from-2010-to-2019/yru6-6re4/data)
* census-data-by-council-district from - [LINK](https://www.kaggle.com/cityofLA/los-angeles-census-data?select=census-data-by-neighborhood-council.csv)
* LAPD\_Police\_Stations from (including information about each station collected from <http://www.lapdonline.org/>) - [LINK](https://geohub.lacity.org/datasets/1dd3271db7bd44f28285041058ac4612_0/data)

Source of general information:

* <http://www.lapdonline.org>
* <https://en.wikipedia.org/wiki/Los_Angeles_Police_Department>
* <https://en.wikipedia.org/wiki/Crime_in_the_United_States>
* <https://www.bjs.gov/index.cfm?ty=tp&tid=3>
* <https://www.justia.com/criminal/offenses>

# Methodology (Project design)

## Data

Here you have to describe how you plan to manipulate the data. For this you have to answer to the following questions:

* Which data will be used?
  + Describe data sources
  + Describe possible external data sources that may enrich our data
  + Data for external validation?
* On which time frames periods will your project will be based on?
  + Time-frame for training
  + Time-frame for test?
* How do you define your subjects?
  + Inclusion criteria?
  + Exclusion criteria?
* Which would be your outcome variable?
* Are there confounder variables that may affect the outcome?
* Is there a possible source of bias in our data?
* Describe your data exploration strategy.
* Which techniques will be applied to enrich the data?
* How will you deal with outliers?
* How you will deal with missing values
* Add at the end of the protocol (appendix) the [Data retrieval protocol](https://docs.google.com/spreadsheets/d/1pYYjgwZ_8PS1Bcmc2kRNHTL0f_rk__GCJALLs1JHPUQ/edit#gid=0)

## Models

Here you have to describe how do you plan to develop your models:

* How do you plan to divide your data
  + Training, validation, test - proportions, techniques
* Do you need to balance your data? How?
* Do you need to stratify/subsample your data? How?
* What techniques will you apply to model your outcome?
  + Unsupervised
  + Regression
  + Classification
* Will you use cross-validation and/or bootstrap?
* Which measures you will use to train and evaluate your models? Why?
* Do you plan to use ensembling or will use your best model?

## Deployment of your model

* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How will the prediction be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* On which platform the predictions will be deployed?
* How frequently will the model be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which models were used and which were selected for the final prediction.
* Which measurements were used to evaluate the prediction.
* Which results we got from those models.

# Results

Here you will present the main results of all the processes. We will describe:

* The final amount of data used (total, train, test, etc)
* The amount of outliers and the way of treating them,
* The amount of missing values and the methods used for imputing them,
* The distribution of the data (timeframes)
* The methods used to transform the data and to generate new features.

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how did you get the final prediction. You have to discuss the limitations of the model, when it can be used and when not.