**Data Science Project Protocol**

*Author: Shiri Margalith*

*Date: Jan 29, 2021*

# 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 and tract characteristics in 2015 aiming to understand which characteristics are correlated with higher crime rate and offer a model to predict the expected crime rate level based on these features.

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. Understanding the tract’s characteristics and how they can predict the crime rate 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 look at various parameters, such as age, ethnicity, education, financial situation and more which will contribute to our understanding of the crime rate. This data is available to us by tracts in Los Angeles, which can then be grouped together for neighborhood analysis.
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 neighborhood in order to better understand the population. This, in the future can be used to develop a taylor-made activity plan for each neighborhood.

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

* Age\_Distribution (count per age group) - <https://usc.data.socrata.com/Los-Angeles/Age-Distribution-LA-/rqg9-k6ju>
* College\_Enrollment (percent) - <https://usc.data.socrata.com/dataset/College-Enrollment-LA-/uxss-3jkp>
* Educational\_Attainment (percent) - <https://usc.data.socrata.com/Los-Angeles/Educational-Attainment-LA-/2qtg-fxmr>
* Employment\_Status (percent) - <https://usc.data.socrata.com/Los-Angeles/Employment-Status-LA-/dxxc-6x7h>
* Homelessness (count) - <https://www.lahsa.org/news?article=406-2015-homeless-count-results%7B&%7Dref=hc>
* Homeownership (percent) - <https://usc.data.socrata.com/Los-Angeles/Homeownership-LA-/misi-374t>
* Immigration (percent) - <https://usc.data.socrata.com/Los-Angeles/Immigration-LA-/bysj-yujw>
* LA\_crimes (count per 1000 people) - <https://usc.data.socrata.com/Los-Angeles/Part-II-Crimes-LA-/hfc5-9p65>
* Marital\_Status (percent) - <https://usc.data.socrata.com/Los-Angeles/Marital-Status-LA-/2i7t-jxzt>
* Median\_Household\_Income - <https://usc.data.socrata.com/dataset/Median-Household-Income-LA-/ani7-k64m>
* Neighborhoods - <https://usc.data.socrata.com/>
* Opportunity\_Youth (percent) - <https://usc.data.socrata.com/Los-Angeles/Opportunity-Youth-LA-/7ayf-d8q3>
* Poverty (percent) - <https://usc.data.socrata.com/Los-Angeles/Poverty-LA-/rrsi-vihu>
* Race & Ethnicity (percent) - <https://usc.data.socrata.com/Los-Angeles/Race-Ethnicity-LA-/jxw5-xxv5>
* Rent – Median rent price per area - <https://usc.data.socrata.com/Los-Angeles/Rent-Price-LA-/4a97-v5tx>
* Total\_Population (count) - <https://usc.data.socrata.com/Los-Angeles/Total-Population-LA-/y65w-5vdw>
* Vehicle Ownership (percent) - <https://usc.data.socrata.com/Los-Angeles/Vehicle-Ownership-LA-/5j9x-k2hp>
* Youth (percent) - <https://usc.data.socrata.com/Los-Angeles/Youth-in-Households-LA-/mvd7-9t72>

Additional sources 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

The data used for this project was gathered from <https://usc.data.socrata.com/> - The Neighborhood Data for Social Change (NDSC) which is a project for Social Innovation. The data is available for civic actors to learn about their neighborhoods, illuminating the trends, challenges, and opportunities facing local communities.

I am using 17 different datasets with various information about the Los Angeles area, that can contribute to our understanding of crime in Los Angeles and predicting it. Combining the various datasets and finding the logic and correlations enables us to make some interesting conclusions, helping the community stakeholders track measurable change and improve local policies and programs.

The analysis will focus on 2015, analyzing the contributor factors which include: population size, age distribution, education, employment, financial indicators (such as poverty, homelessness, income, vehicle ownership and homeownership), immigration, race & ethnicity, marital status, youth in household and opportunity youth.

The LA area has 226 neighborhoods, each one is divided into several tracts, so in total we have 1627 tracts to analyze.

Inclusion criteria – all tracts included in this project have information about the crime rate in 2015. They also have information on over 30% of the features. tracts that are lacking number of crimes in 2015 or have less than 30% of the features will be excluded.

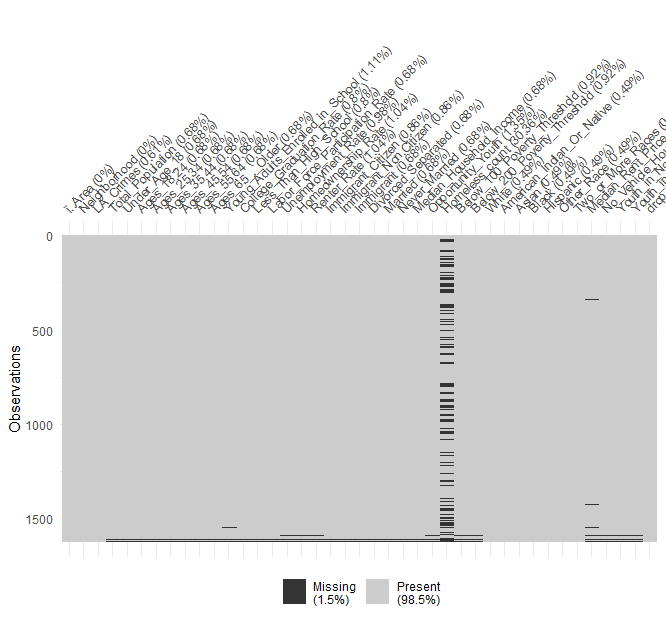
Our outcome variable is the number of crimes per 1000 people, based on the predictive features.

**Data exploration with r:**

1. Flat file creation: The data was extracted from the datasets in the above links. The flat file was created in SQL, joining all datasets by the tract number into one table, providing all information for each tract.
2. File preparation: The tract and neighborhood are characters while all other features are numeric.
3. Data exploration:
   1. First step will be looking at the target variable. I see that it has some significant outliers and also that it has some missing values. The rows with missing values at the target variable will be excluded from the project.
   2. Looking at the data, I grouped the information by neighborhood and listed the top 5 neighborhoods with highest crime rate per 1000 people:
      1. Griffith Park
      2. Palmdale
      3. Chatsworth Reservoir
      4. Downtown
      5. Van Nuys
   3. Next is descriptive statistics, using ExploreData. The information can be found in two files:
      1. **1 Data Visualization ORIGINAL DATA**
      2. CrimeData - Data Retrieval Protocol
   4. Looking at the data, there are significant outliers that impact the distribution of the target variable but I will first look at rows that have a lot of missing values:

The dataset has 1617 rows, out of which 1112 (68.8%) are complete.

The Largest number of missing is at Homeless\_count – 486 missings (30.1%). The next one is Median\_Rent\_Price with only 32 missings (2%). This can be clearly seen here in the missing matrix:



Only 5 rows have 30% or more missing. We can drop them. Since it is a very low number (out of 1617 rows) I am using 30% as the cutoff, leaving the table with 1612 rows.

After removing the missing – 1112 rows (69%) are complete out of 1612 rows.

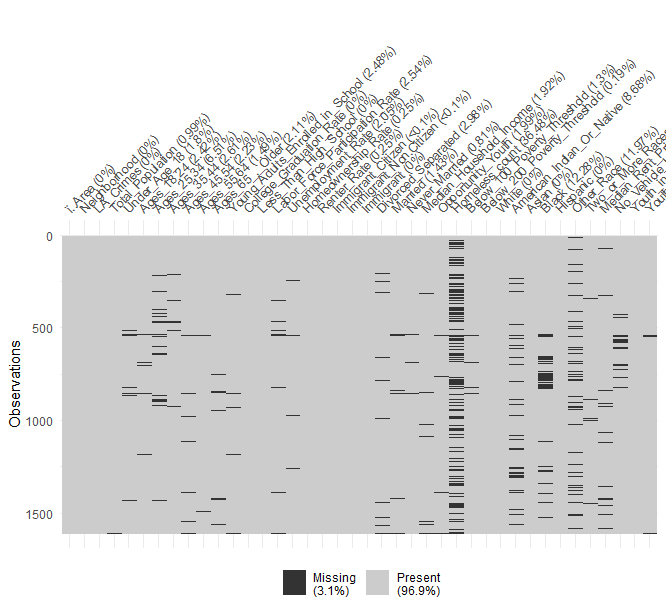
Only 14 features have missing items now. Homeless\_count is still with most missings: 481 (29.8%), followed by Median\_Rent\_Price – 27 (1.7%).

Using KOLMOGOROV-SMIRNOV test, as well as GLM, we see that there is no impact of the missing, they are missing at random and we can make imputation on all of them. But we will first check the outliers.

We see that all outliers can be dropped except for LA\_Crimes (our target), Asian, and Youth\_in\_Non.Family\_Households.

After dropping the outliers I will look at the missing again. Now there are only 568 complete rows (35.2%) out of 1612. Most of them are at Homeless\_count (588 – 36.5%) followed by Black (198 – 12.3%). Median\_Rent\_Price has dropped lower now.

Here is the updated missing matrix after dropping the outliers:



Now we have 8 rows with 30% or more missing – I will drop them as well. We will have 1604 rows left with 568 complete rows (35.4%).

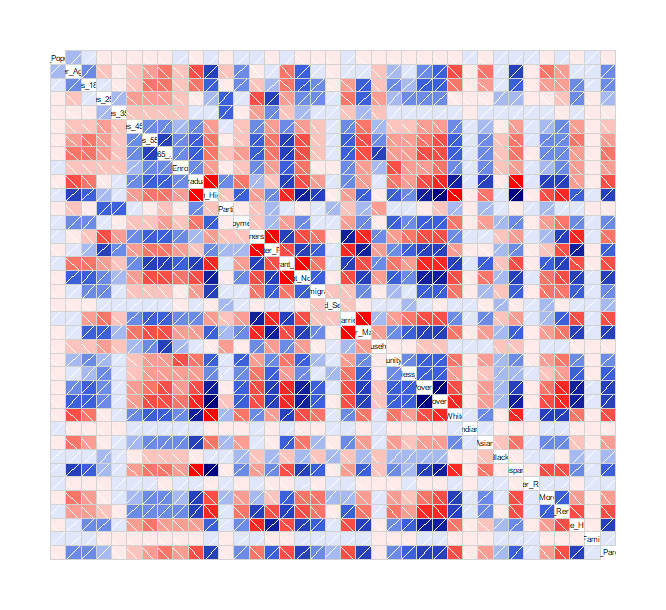
Again, using KOLMOGOROV-SMIRNOV we see that there is no impact of the missing and we can make imputation to all of them.

1. I will use KNN to for imputation and will create the full table: Crimes\_data\_full.
2. I will now look at the correlation of each feature with the target, both visually and statistically, comparing the correlation. The visual data can be found at **4 Data Visualization FULL**. The following table shows the features whose correlation with the target is strongest:



1. In order to enrich the data I compared the variables and found several cases in which two features had strong correlation with one another. Data enrichment was then performed, creating new features from existing ones with high correlation.

The correlation among the original features can be seen in the following image and the table that follows, reflecting the strongest correlation between features.





Based on these findings I have decided to create the following new features:

* + 1. YoungVsOld: The ratio of young population (24 and under) and older population (55 and older)
    2. OpportunityVsYoung: The ratio of opportunity youth from the entire young population (24 and under)
    3. EnrolledVsYoung: The ratio of young population enrolled in college from the entire young population (24 and under)
    4. NoHighSchoolVsCollege: the ratio of those with no high school deploma from those who graduated from college
    5. Singles: The number of those who are separated, divorced and never married
    6. NonWhite: The number of all ethnic groups that are not white
    7. WhiteVsNonWhite: The ratio of white population from the non-white population
  1. After the data enrichment I have 47 features and I checked the correlation between the features. The following have high correlation and can be dropped so we are left with 30 features. The following are the ones I removed:
     1. Homeless\_count
     2. Renter\_Rate
     3. Homeownership\_Rate
     4. Opportunity\_Youth
     5. Young\_Adults\_Enrolled\_in\_School
     6. Divorced\_Separated
     7. Asian
     8. EnrolledVsYoung
     9. Below\_100\_Poverty\_Threshold
     10. Ages\_65\_.\_Older
     11. Immigrant
     12. Ages\_45.54
     13. Labor\_Force\_Participation\_Rate
     14. Unemployment\_Rate
     15. NoHighSchoolVsCollege
     16. YoungVsOld
     17. Never\_Married

## Models

I ran the models, splitting the data into 3 parts: train, dev and test. Since I have ~1600 rows, I allocated 20% to the test and out of the rest – 20% to dev and the rest to train, making sure I have perfectly balanced datasets.

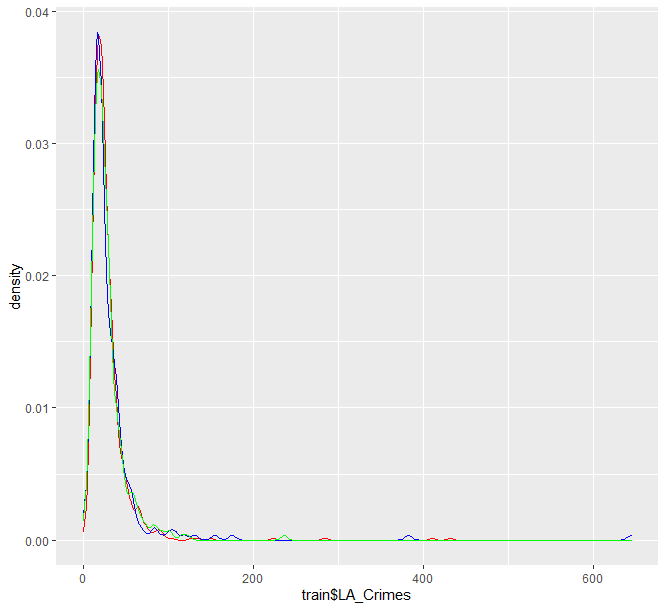
Entire dataset – 1604 rows

“test” – 321 rows (20%)

“train” – 1026 rows (80% of the remaining 1283)

“dev” – 257 rows (20% of the remaining 1283)

The following graph shows that the split is balanced between the three datasets:



I will try various models using the RMSLE metric for comparison. I will then fine tune the best model. Out of the model checked (linear regression, tree, random forest, ranger, knn, SVM and XGBoost) XGBoost provided the best result and then I fine tuned the parameters (max\_depth, min\_child\_weight, subsample, nrounds and eta).

I then tested the final improved model on the test file and received 98% accuracy.

Based on this model, we can predict the crime rate in various areas of Los Angeles based on the features we used. As described in the introduction, these result will help the LAPD plan its workforce across the city in order to avoid crimes as much as possible and to quickly react to them as they occur. This model can be updated on a yearly basis, as the demographic and socio-economic data of the city is updated.

# Deployment and Results

The model predicts the number of crimes per 1000 people based on the different features. As the population changes these features may change. In addition, with proper social care and police treatment, the number of crimes in the riskier areas may drop. Therefore, I suggest running the model on yearly basis, including the most recent data changes.

When running the model it is important to keep in mind the following points:

1. The merge of the tables in SKL is done on the column named “Area” that has the tract number of each item.
2. All features were numeric and continuous (except for the Neighborhoods’ names and the Areas’ numbers, which are not used for prediction). The model is based on all features being as such, so if for any reason the information is changed and a feature because categoric it needs to be updated.

The outcome of this project is aimed to help the LAPD prepare to fight crime and also provide the needed support to various populations in the different neighborhoods.

It collects the demographic, socio-economic and ethnical data in one place, showing the correlation between the different features, and enabling prediction per tract. The usage of tracts leads to more accuracy but I would suggest that the findings will be dealt with per neighborhood for better allocation of the LAPD force.

The model will present the top 10 neighborhoods with the highest expected crime rate per 1000 people as well as the neighborhood’s other characteristics in order to help the social services in Los Angeles as well.

To summarize the process we started with a dataset of 1627 rows, each one representing a tract in Los Angeles. After removing rows with no crime data and with 30% and over of missing values we were left with 1604 rows. We cleared the outliers where it did not impact the distribution and replaced them, as well as filling the missings using knn imputation.

The detailed information of the features used can be found at the data retrieval protocol attached.

# Conclusion

The streets of Los Angeles always seemed frightening and I was curious about the potential crime rate in each neighborhood. This project was an opportunity to understand the differences between the parts of the city, and how those differences impact the potential crime rate. Although a strong correlation was found between the target (crime rate) and several features, looking at all of them as a whole provided a much better projection of potential crime.

The main difficulty was collecting the information from various datasets. Furthermore, the latest data I found that had all features was from 2015. It’s been five years since then, and I assume COVID will also have an impact on the results. I would consider updating the model and checking the projection on 2020 data when it is available.