ISyE 6740 – Spring 2021
Project Proposal

Project Group: C
Team Number: 85
Team Member Names: Yang Xu, Liu Shizhu, He Xiao
Yang Xu GTID: 903-657-847
Liu Shizhu GTID: 903-540-469
He Xiao GTID: 903-560-741

Project Title: Houston Crime Data Analysis
07/15/2021

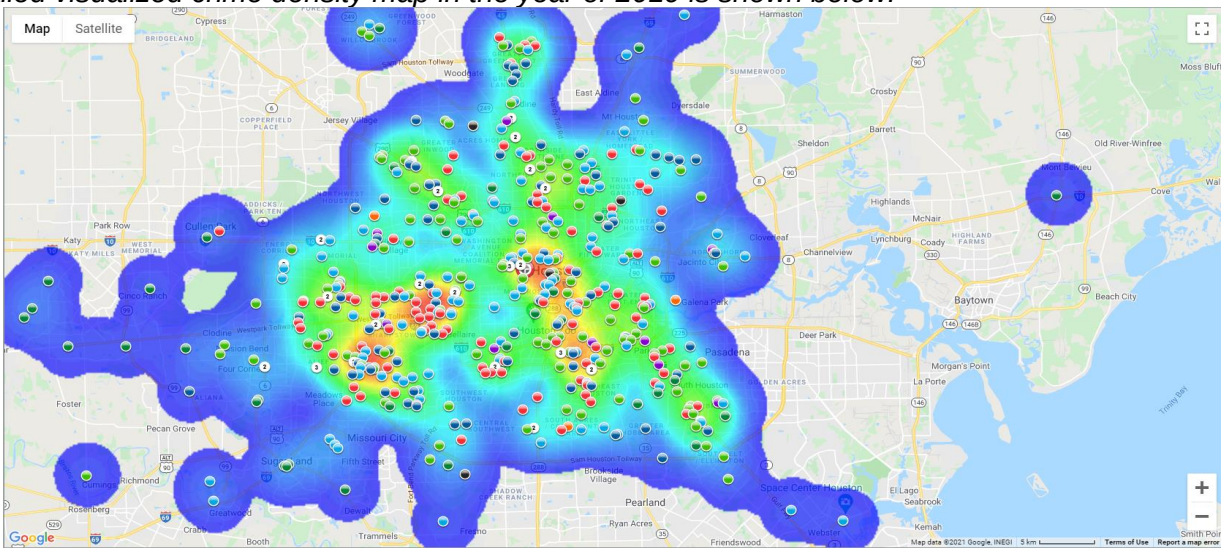
Table of Content

1. Problem Statement

Houston is the most populous city in the U.S. state of Texas, fourth-most populous city in the United States, most populous city in the Southern United States, as well as the sixth-most populous in North America, with an estimated population of 2,320,268 in 2019. Located in Southeast Texas near Galveston Bay and the Gulf of Mexico, it is the seat of Harris County and the principal city of the Greater Houston metropolitan area, which is the fifth-most populous metropolitan statistical area in the United States. Houston is the southeast anchor of the greater megaregion known as the Texas Triangle.[1]

However, the famous city is suffering from crime and gangs' problems. Across the Houston city, crime incidents were up 13% in 2020, with 256,833 crimes reported to Houston Police Department in 2020 compared to about 227,000 crimes reported in 2019.[2]

A detailed visualized crime density map in the year of 2019 is shown below:[3]



In this project, we are trying to use both supervised and unsupervised machine learning algorithms, to get some insights on following questions, and hopefully provide the Houston Police Department some prescriptive measures to lower the crime statistics in the future:

1. Where in Houston is relatively facing more crime? What is the trend? (Geographic)
2. What time in Houston is relatively facing more crime? What is the trend? (Time series)
3. What type of crime happens more frequently? What is the trend? (Descriptive)
4. Is there any correlation between crime rate and local unemployment rate? (Prediction)
5. Is there any correlation between crime rate and local poverty? (Prediction)
6. Is there any correlation between crime rate and local income? (Prediction)
7. Is there any correlation between crime rate and local education level? (Prediction)
8. Is there any correlation between crime rate and weather? (Prediction)
9. How will the crime rate influence local housing prices? (Prediction)
10. How can we use these insights to guide local police resources? (Prescription)
11. How can we use these insights to guide people who visit Houston? (Prescription)
12. If a successful crime is carried out, will it happen again within 1 month? (Bayesian Inference)

[1] Introduction of Houston City, https://en.wikipedia.org/wiki/Houston

[2] 13 Investigates: Crime reported every 7 hours in this Houston neighborhood, ABC news, https://abc13.com/houston-crime-increasing-on-the-rise-neighborhoods-with-most-police-department/10560344/

[3] Community Crime Map, https://communitycrimemap.com/#nwrButton

2. Data Source

As a data analytics project, our main goal relies on reliable and trustworthy data sources. Therefore, after brainstorming and consulting with criminal justice/political science domain experts, we collected our data source as follows:

We download the Houston crime data from City of Houston Police Department (https://www.houstontx.gov/police/cs/Monthly_Crime_Data_by_Street_and_Police_Beat.htm), with 219034 rows of data points, 14 columns of features.

We download the unemployment data from BLS Beta Labs (https://beta.bls.gov/dataViewer/view/timeseries/LASST480000000000004), with 12 rows of data points for each month, 5 columns of features.

We download the median household income data from Houston State of Health (http://www.houstonstateofhealth.com/indicators/index/view?indicatorId=315&localeTypeId=3), with 1165 rows of data points, 12 columns of features.

We download the Houston poverty data from Houston State of Health (http://www.houstonstateofhealth.com/indicators/index/view?indicatorId=240&localeTypeId=3), with 1171 rows of data points, 16 columns of features.

We download the Houston education data from Houston State of Health (http://www.houstonstateofhealth.com/indicators/index/view?indicatorId=341&localeTypeId=3), with 1172 rows of data points, 15 columns of features.

We download the Houston housing data from zillow data research engine (https://www.zillow.com/research/data/), with 30837 rows of data points, 314 columns of features.

We download the Houston weather data from visualcrossing (https://www.visualcrossing.com/weather/weather-data-services#/login), with 17522 rows of data points, 17 columns of features.

After cleaning the data, we use zip code and crime occurrence time (e.g., year, month, day, and hour) to combine these data for our analysis.

3. Methodology

For the purpose of crime prediction, we plan to implement algorithms which learned in the lesson, based upon the actual analytical needs.

1) **Clustering.** Based upon this unsupervised learning algorithm, we plan to gain some insights about the clusters of crime statistics and hope to compare with the NIBRS reporting system label. While we don't know exactly how many clusters there should be or what each cluster means, this type of analytical process could give us a sufficient idea of the incidents that happened in the year of 2019. Among all clustering methods, we plan to compare hard clustering label method K-means and probability-based modelling GMM.

2) **Regression**. Our main goal is to provide a solid prediction in the future to give the Houston Police Department a reasonable and legal practice for future affairs. Therefore, our main focus of modelling this project will be on prediction using regression. While we collected lots of feature variables candidates, we will first perform dimension reduction through PCA and then build regression models upon projected data. LASSO and RIDGE regression methods will also be implemented to make sure our analysis does not involve too many factors which are not easy to interpret.
3) **Classification.** We want to use decision-tree-based ensemble algorithms to classify the criminal cases and evaluate the model based on the NIBRS reporting system label. By using this analytical method, we would get to know which features or crime characteristics are important in predicting crime categories of Houston. We plan to compare the results from random forest, gradient boosting, and xgboost.

4. Evaluation and Final Results

We will pick the most 5 convincing results when comparing different algorithms.
For clustering, we will use accuracy as the judgement of each algorithm.
For regression, we will use root mean squared error (RMSE) as the judgement of each algorithm.
For classification, we will use F-1 score as the judgement of each algorithm.

Besides the quantitative analysis, we will also consult & communicate with criminal justice/political science domain experts for qualitative analysis. (Already communicated with several master students and contacted Houston Police Department - Community Affairs Department).

We will finish this part in the final project report.

5. Untechnical questions

When we are analyzing the data, besides the technical methodology, we also need to make sure that our analysis follows the high ethical standards.

We are dropping the following categories of discrimination data:
Sex or gender identity
Mental disability or physical disability
Religion
National origin or ethnicity
Sexual orientation
Marital, veteran, and parental status

Besides these, when we get the insights of the crime data and reallocate our resources, we also need to investigate the impact to local communities and businesses. For example, when we have more patrols in a specific community, it will bring a negative impact on their local stores and residents. We should balance the insights and impact for lowering the crime rate step by step and not disturbing the community too much.

It also needs to be mentioned that although the algorithm may make our model reach to 99.99% or even 100% accuracy, we should never make our judgement purely depends on prediction. When solving the crime related question, the poor suspect may be the 0.01%, and the judgment for him will destroy his or her future.

6. The future work:

Due to limited time, this project discussed the following parts:
Will added in the final report.

The future work followed by our report may include:
Will added in the final report.

7. Appendix

Will added in the final report.