ISyE 6740 – Spring 2021
Final Report

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

08/02/2021

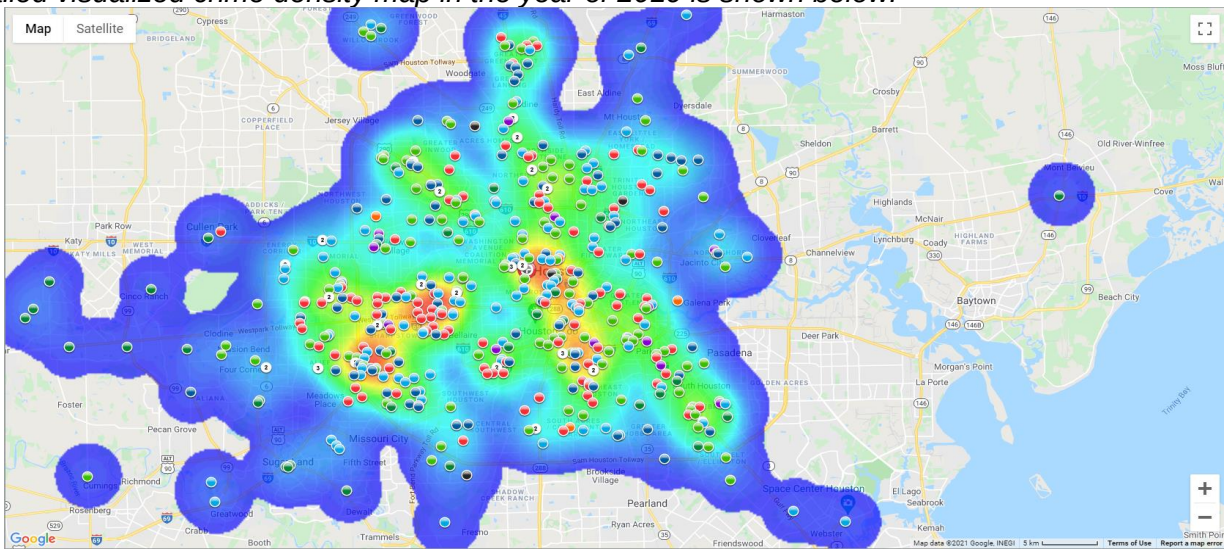
Table of Content

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 provide the Houston Police Department some prescriptive measures to lower the crime statistics in the future:

1. Where in Houston Chinatown is relatively facing more crime?
2. What time in Houston is relatively facing more crime?
3. Which month in Houston is relatively facing more crime?
4. Is there any correlation between crime rate and local unemployment rate?
5. Is there any correlation between crime rate and local poverty?
6. Is there any correlation between crime rate and local income?
7. Is there any correlation between crime rate and local education level?
8. Is there any correlation between crime rate and weather?
9. How can we use these insights to guide local police resources?

[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 implemented algorithms which were learned in the class, based upon the actual analytical needs. In order to fit the local police station's needs, besides the initial data density estimation, we shrunk our data into the most crime incidents' zip code "77036" which is Houston Chinatown.

3.1 Clustering

Based upon this unsupervised learning algorithm, we gained 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 compared the hard clustering label method K-means and K-modes.

3.2 Classification

We used all algorithms that we learned in the class such as 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 are important in predicting crime characteristics of Houston. We compared the results from all algorithms and picked the most convincing one to generate more insights.

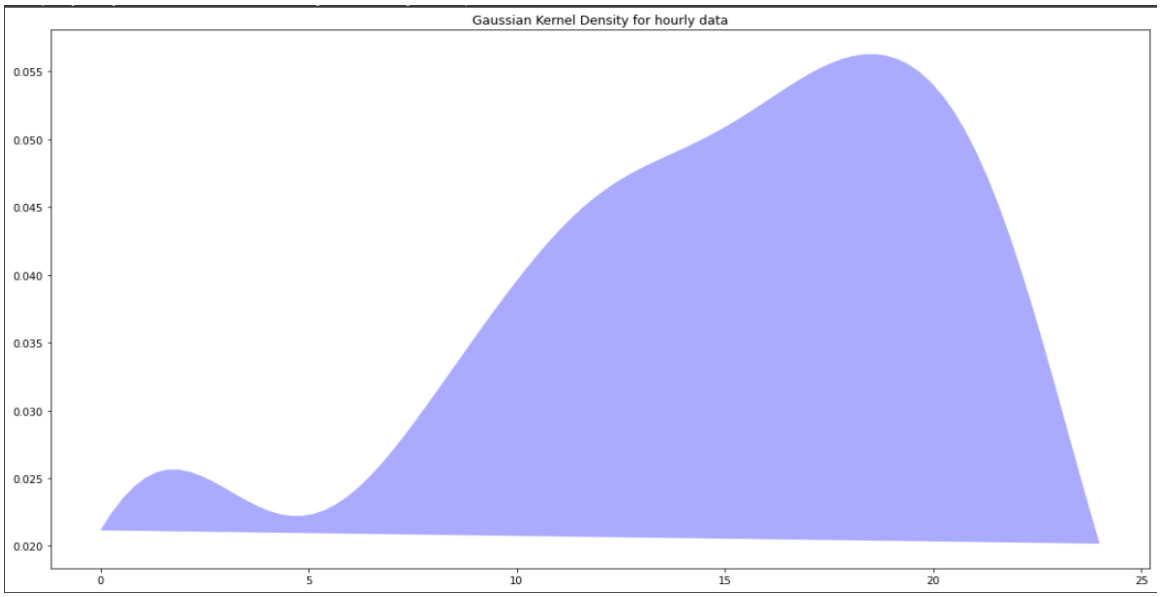
3.3 Regression

Our main goal is to provide a prediction in the future to give the Houston Police Department a reasonable and legal practice for future affairs. Therefore, one of modelling this project will be on prediction using regression. While we collected lots of feature variables candidates, LASSO regression methods will be implemented to make sure our analysis does not involve too many factors which are not easy to interpret.

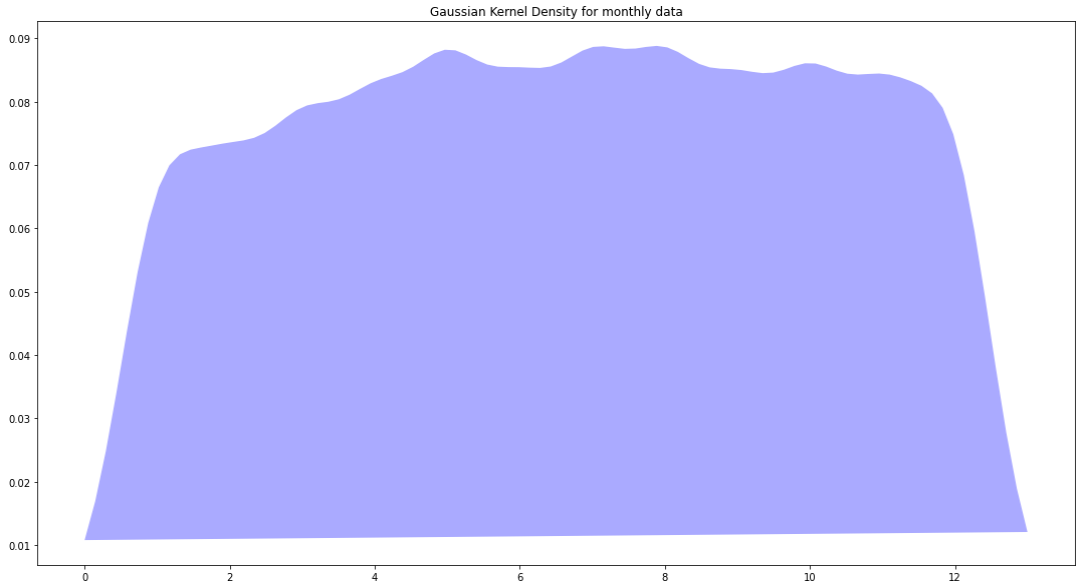
4. Evaluation and Final Results

4.1 Data Density Estimation

First of all, we used KDE to draw the total crime numbers on different hours and different months, based on crime data of the year of 2019, so that we could gain some idea about what's the estimate distribution of these two factors:



As can be seen from the Kernel Density for hourly data, we could observe that most crime is occurred between 17:00 to 22:00. During 0:00 to 8:00, the crime will relatively be observed less.
This could lead to the guidance that for the Houston Police department, more patrols could be arranged during evening time.



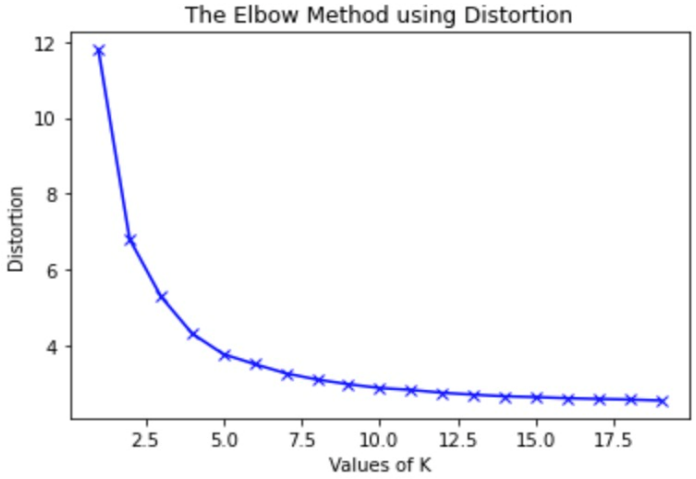
As can be seen from the Kernel Density for monthly data, we could observe that most crime is occurred in May, July and August. During January and February, the crime will relatively be observed less. This could lead to the guidance that for the Houston Police department, more police power could be distributed during Summer and Autumn "busy" sessions.
Therefore, having obtained the initial density information about the crime temporal information, we should be able to perform analysis afterwards.

4.2 Clustering

In order to fit the local police station's needs, we shrunk our data into zip code "77036" in which the most crime incidents occurred in the Greater Houston area.

4.2.1 K-means clustering on 7 features

When we use K-means clustering, we choose 7 features, including "Hour" (incident hour), "Category" (case category), "Beat" (patrol area represented by 5-digit), "Premise" (incident location), "StreetName" (incident street), "Month" (incident month), and "Weather_conditions" (weather data when incident happened). We used one-hot encoding to convert the 7 categorical variables.
To better decide the value of K, we use elbow method to select optimal K values:



As can be seen in the plot, we should use K=5 to cluster the data in order to avoid overfitting.

By deploying the K-means clustering with K=5, we get 5 cluster centers (shown as below):

	Hour	Category	Beat	Premise	StreeName	Month	Weather_condition
0	20	Theft from motor vehicle	19G10	Residence, Home (Includes Apartment)	BISSONNET	3	Clear
1	22	Simple assault	18F60	Residence, Home (Includes Apartment)	BELLAIRE	7	Clear
2	20	Destruction, damage, vandalism	19G10	Residence, Home (Includes Apartment)	BELLAIRE	1	Clear
3	20	Simple assault	18F60	Residence, Home (Includes Apartment)	BELLAIRE	12	Clear
4	12	Theft from motor vehicle	18F60	Residence, Home (Includes Apartment)	BELLAIRE	8	Clear

It can be seen, for zip code 77036, the 5 most representative crimes are all occurred in residence when the weather is clear, with patrol code “19G10” and “18F60” at Bissonnet Street and Bellaire Street, in January, March, July, August, and December, at 12pm, 8pm and 10 pm, and categorized as “Theft from motor vehicle”, “Simple assault” and “Destruction, damage, vandalism”. This could be adding a lot of value to the local HPD’s practices.

4.2.2 K-means clustering on 3 features

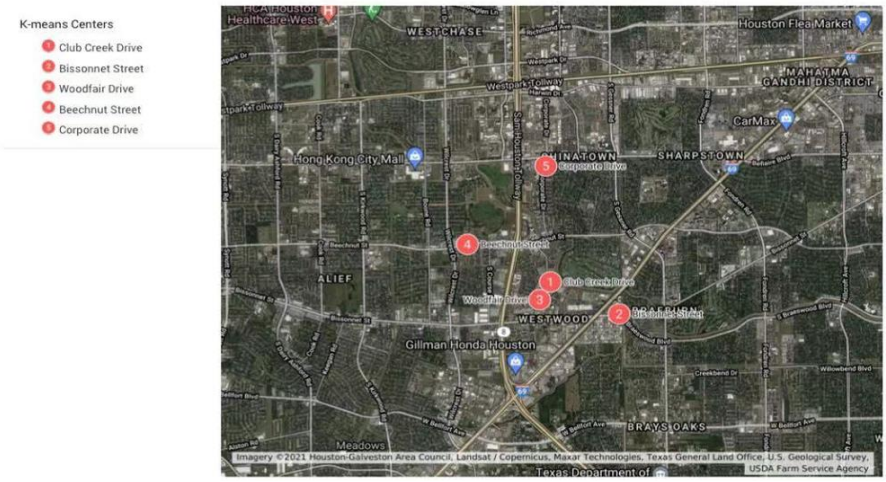
When we communicated with domain expertise from the HPD community affairs officer, they asked us if we could use our models to give them more guidance on when and where to patrol based on their working schedule.
We then used K-means clustering on only “Hour”, “Premise”, and “StreetName”. In order to fit their working schedule, we filtered Hour between 14:00 to 22:00, and focused on the Beat=19G10 area. We used one-hot encoding to convert the 3 categorical variables to numeric so that it can be fit into the clustering algorithm.

Based on local police officers’ patrol time (i.e., 14:00 to 22:00), we choose K=8, and here are the 8 clustering centers (shown in the table below):

	Hour	StreetName	Premise
0	22	CLUB CREEK	Highway, Road, Street, Alley
1	14	WOODFAIR	Residence, Home (Includes Apartment)
2	20	CLUB CREEK	Parking Lot, Garage
3	18	BEECHNUT	Restaurant
4	20	CLUB CREEK	Residence, Home (Includes Apartment)
5	17	BISSONNET	Parking Lot, Garage
6	22	CORPORATE	Residence, Home (Includes Apartment)
7	16	BISSONNET	Highway, Road, Street, Alley

We also pinned these street names on Google map:

Kmeans map



It seems that the cluster center streets are distributed across the main streets that behave like surrounding the zip code 77036.

4.2.3 K-modes clustering on 3 features

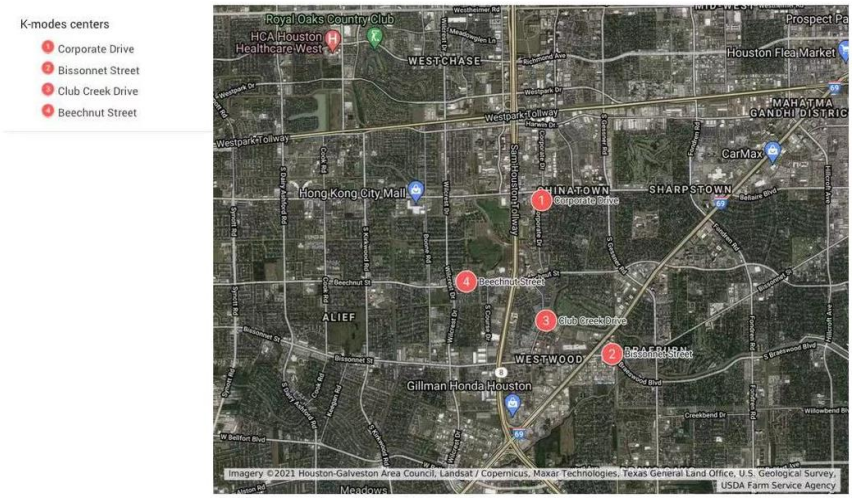
Besides K-means clustering, we also used K-modes clustering on “Hour”, “Premise”, and “StreetName” to compare the differences. We filtered Hour between 14:00 to 22:00, and filtered Beat districts only in the 19G10 area as well.

We used K=8 based on local police officers’ patrol time, and here are the 8 clustering centers (shown in the picture below):

```
['22' 'BISSONNET' 'Highway, Road, Street, Alley']
['21' 'CLUB CREEK' 'Residence, Home (Includes Apartment)']
['20' 'CORPORATE' 'Residence, Home (Includes Apartment)']
['21' 'CORPORATE' 'Other, Unknown']
['14' 'BEECHNUT' 'Parking Lot, Garage']
['18' 'BISSONNET' 'Highway, Road, Street, Alley']
['20' 'CLUB CREEK' 'Residence, Home (Includes Apartment)']
['15' 'CORPORATE' 'Residence, Home (Includes Apartment)']]
```

Similarly, we pinned these street name on Google map:

Kmodes map



Thus, we can see that the results from K-means and K-modes are similar to each other. Based on the results from the two models, Beechnut Street, Bissonnet Street, Club Creek Drive, and Corporate Drive can present the street locations of all the clusters. As we can see from the Google map, the result makes sense in that these streets are the several main streets of this area.

However, because K-means and K-modes are unsupervised learning methods, we cannot use metrics to evaluate their performance. We might get more feedback from domain experts on the actual results of using these two different maps and one random patrol map, and see if there is mathematical significance on different patrol teams.
We will mention more insights in 4.4 Domain Expert Consultation.

4.3 Classification

Other than the unsupervised learning, which was mainly clustering and density estimation, we also plan to gain some insights from the data to see if we could predict the outcome category, premise or street name using the data features we had. For classification, because we are trying to predict the crime, and use the insights to guide the patrol, the cost of False Negative (a crime is happening but we think it is not) is much higher than False Positive (we think the crime is happening but we are wrong).
So, we will use accuracy and recall score as the performance metrics of each algorithm.

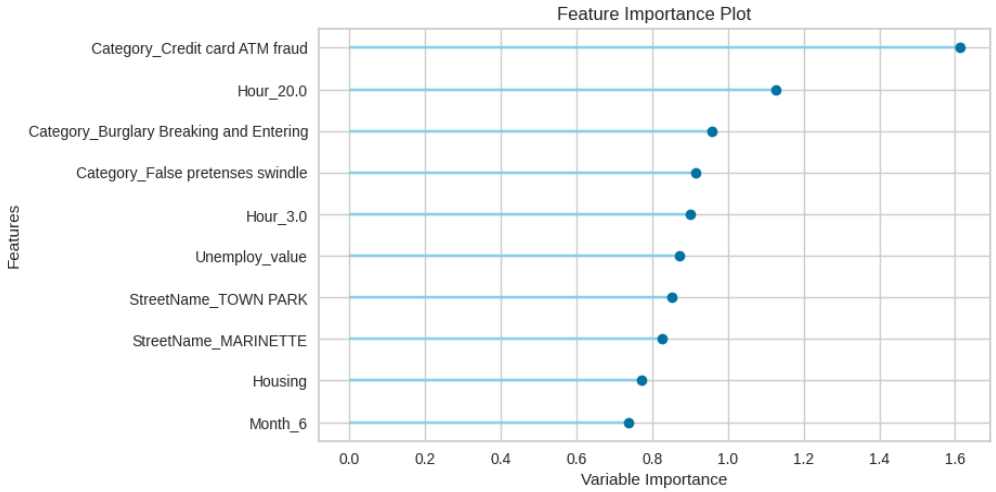
4.3.1 Predict "Premise"

When we are using "Hour", "Category", "Month", "Weather_conditions", "Housing", "Temperature", and "Unemply_value" to predict the "Premise", we used Logistic Regression, Random Forest Classifier, Ridge Classifier, SVM - Linear Kernel, Decision Tree Classifier, K Neighbours Classifier, AdaBoost Classifier, and Naive Bayes as the candidate algorithms. We ranked them from highest to lowest based on Accuracy and Recall score:

Model name	Accuracy	Recall
Logistic Regression	0.5744	0.1673
Random Forest Classifier	0.5635	0.1543
Ridge Classifier	0.5627	0.1023
SVM - Linear Kernel	0.5539	0.1794
Decision Tree Classifier	0.4629	0.1563
K Neighbours Classifier	0.4145	0.1116
AdaBoost Classifier	0.4098	0.0455
Naive Bayes	0.0897	0.1279

In addition, for Logistic Regression, we checked the 10-fold validation accuracy and recall score, and then plot the important features:

Fold	Accuracy	Recall
0	0.5520	0.1606
1	0.5978	0.2099
2	0.5368	0.1397
3	0.5925	0.1641
4	0.5512	0.1912
5	0.5709	0.1971
6	0.6086	0.1727
7	0.5745	0.1330
8	0.5799	0.1559
9	0.5799	0.1484
Mean	0.5744	0.1673
SD	0.0214	0.0240



From the above feature importance plot, we drew the conclusion that in order to predict the correct crime premise category, the credit card ATM fraud and the Hour 20:00 contributes the most, meaning we could notify the police of potential crime premises based on these, however, it seems dumb to say that wherever is a credit card fraud, the crime will happen at the specific premise. We plan to dig into this further after discussion with domain experts.

4.3.2 Predict "Hour" category

When we are using "Premise", "Category", "Month", "Weather_conditions", "Housing", "Temperature", and "Unemply_value" to predict the "Hour", we used Logistic Regression, Random Forest Classifier, Ridge Classifier, SVM - Linear Kernel, Decision Tree Classifier, K Neighbours Classifier, AdaBoost Classifier, and Naive Bayes as the algorithms. We ranked them from highest to lowest based on Accuracy and Recall score:

Model name	Accuracy	Recall
Decision Tree Classifier	0.3026	0.2837
Random Forest Classifier	0.2237	0.2069
K Neighbours Classifier	0.1328	0.1258
Logistic Regression	0.1081	0.0923
Ridge Classifier	0.1021	0.0831
AdaBoost Classifier	0.0860	0.0732
SVM - Linear Kernel	0.0795	0.0738
Naive Bayes	0.0278	0.0493

As can be seen, unfortunately, there are no good algorithms to reach an acceptable accuracy and recall score to predict crime hour.

4.3.3 Predict "Category"

When we are using "Premise", "Hour", "Month", "Weather_conditions", "Housing", "Temperature", and "Unemply_value" to predict the "Category", we used Logistic Regression, Random Forest Classifier, Ridge Classifier, SVM - Linear Kernel, Decision Tree Classifier, K Neighbours Classifier, AdaBoost Classifier, and Naive Bayes as the algorithms. We ranked them from highest to lowest based on Accuracy and Recall score:

Model name	Accuracy	Recall
Ridge Classifier	0.2608	0.1241
Logistic Regression	0.2587	0.1272
Random Forest Classifier	0.2314	0.1189
SVM - Linear Kernel	0.1971	0.1034
AdaBoost Classifier	0.1764	0.0434
Decision Tree Classifier	0.1736	0.0956

K Neighbours Classifier	0.1619	0.0856
Naive Bayes	0.0433	0.0703

As can be seen, unfortunately, again, there are no good algorithms to reach an acceptable accuracy and recall score to predict the crime category.

4.4 Regression

Even though for classification we did not obtain sufficient accuracies, we did not give up on supervised learning algorithms. Therefore, we would like to see if a regression model would be appropriate for correlating the crime totals using our collected data features. In this scenario, we did not focus on only 77036 zip code, instead, we aggregated the whole data frame by each zip code and performed the analysis using that information obtained.

For regression, we will use mean absolute error (MAE) , and R2 as the judgement of each algorithm. We are using “Poverty_rate”, “Median_household_income”, “25+_with_high_school_or_higher”, “Average_temperature”, “Average_unemploy_rate”, “Average_housing_price”, to regress the “total_crime_by_zip”. We standardized all the numerical features. We compared the performance between Random Forest Regressor, K Neighbours Regressor, AdaBoost Regressor, Elastic Net, Ridge Regression, Lasso Regression, Linear Regression, and Decision Tree Regressor models. After rigorous parameter tuning, we ranked them from highest to lowest based on MAE, and R2 score:

Algorithm	**MAE**	**R2**
Random Forest Regressor	700,7746	0.6631
K Neighbours Regressor	856,6285	0.3099
AdaBoost Regressor	962,7724	0.2159
Elastic Net	1135,0491	0.1957
Ridge Regression	1177,4625	0.1194
Lasso Regression	1079,1790	0.1048
Linear Regression	1079,7121	0.1025
Decision Tree Regressor	1275,6379	-0.6981

Random forest regressor was chosen to be our optimum model based on the above table. For this model, after standardizing the data, we checked the importance of each feature (shown as in the table below):

	Poverty_rate	Median_household_income	25+_w_high_school_or_higher_in_percent	Temperature_y	Unemploy_value_y	Housing_y
0	0.195648	0.295845	0.08568	0.162563	0.206675	0.05359

Based on the numbers, we could say that the most three relevant data is median household income, average unemployment rate, and poverty rate to predict crime numbers. As a consultant from a domain

expert, this is what they use daily as well to predict where to focus on when they have limited human resources.

4.5 Domain Expert Consultation

Besides quantitative analysis, we also consulted with criminal justice/political science domain experts for qualitative analysis.
We have talked with the Houston Police Department - Community Affairs Department and students living in Houston 77036.
It is proved by the police department that significant crimes occurred in the most busy street: BISSONNET street, which is identified by both K-means (7 features and 3 features) and K-modes clustering(3 features).
We also consulted students living in Chinatown and they diagnosed a lot of bars and restaurant in our Kmeans map and Kmodes map. It is good to know areas where we identified have a bad reputation on safety.

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 allocate 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 (e.g., BISSONNET street), 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.

6. The future work

Due to limited time, this project discussed the following parts:
in zip code 77036
filtered Beat only in 19G10 area due to response of local police department

The future work followed by our report may include:

1) more zip code analysis
2) collect more years' data to see the trend of crime by leveraging time series analysis
3) Develop more reliable classification models using most categorical features

7. Group Work

Yang Xu	Collected the feature data from public resources (including temperature, housing, poverty rate) and response vector from HPD (publised), transformed features and aggregated necessary features and performed feature engineering, developed the main structure of modeling codes in python and model hyperparameter tuning, provided feedback on the final reports.
Liu Shizhu	Searched online public data regarding education, unemployment, and household income and merged them with the dataset created by Yang Xu using zipcode and time; developed some codes for feature engineering and modeling; created the pinned Google Map; provided feedback on the final report.
He Xiao	Project Management like meeting arrangement, meeting minutes, schedules; Wrote project proposal and final report in first version; Initialed project case scenarios and brainstorming with team for more ideas.