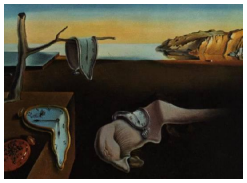


Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Team 13: Chicago Crime Analysis in the Project Final Report Stage

Yu Wang (yw78)
UIUC

Yijia Hu (yijiahu3)
UIUC

Abstract

This project aims to analyze dynamic criminal data in the city of Chicago with explanatory data analysis and model prediction. Our task was to provide a comprehensive analysis of the crimes that took place in Chicago starting from January 1, 2020. The central idea of the project is constructing a complete and robust pipeline to retrieve dynamic data using API, store, provide explantory data analysis with visuals, and conduct predictive analysis on whether an arrest will be made. We imported data with a robust framework, and used data visualization tools to identify crime patterns and distributions and to describe historical trends, insights. In addition, we used logistic regression for training and crime prediction. Discussions on future investigation can also be found. The proposed model has an accuracy of 87.3%.

Keywords: R, group project.

1. Introduction

The City of Chicago has long struggled with high crime rates, and developing effective crime prevention strategies has proven to be a difficult task, especially since the COVID-19. This project aims to use data analytics techniques to analyze crime data in Chicago and identify crime patterns.

The project has two main objectives. First, we want to construct a complete and stable pipeline using dynamic data, considering most up-to-date paper or websites using historical data that lack of timeliness. Secondly, we will figure out crime distributions and patterns with exploratory data analysis and predictive analysis, to provide some insights for the purpose of crime governance and prevention with.

Our motivation for pursuing this project is to contribute to the development of effective

crime prevention strategies in Chicago. We hope to gain a deeper understanding of patterns behind the criminal cases, and we believe effective analysis from organized data helps direct prevention efforts to the safety of Chicago.

2. Related Work

In recent years, crime analysis has become an important field for researchers and policymakers alike. One of the challenges in crime analysis is the large volume of data that must be processed and analyzed. In this context, using R language to analyze crime data is a popular method, as they provide powerful tools for data manipulation and analysis.

The paper *Chicago Crime Analysis using R Programming* by Monish N Monish (2019) focuses on the analysis of crime data in the city of Chicago using R. The author demonstrates the use of various visualization like heat map, and statistical techniques such as K-Nearest Neighbor (KNN) classification to explore crime patterns in the city. The way to choose explanatory variables for prediction based on the pre-processing done during exploratory in this paper is highly appreciated and used for our reference. However, in our project, we applied a different statistical model for prediction with higher accuracy.

Predictive Policing in Crime analysis using R Mallula and Chowdary (2018) focuses on the use of predictive modeling techniques to to distinguish or identify potential criminal activity. Here the paper focused on extracting data directly from the web and then handle the data (which involves cleaning and re-organizing and processing the crime record data), recover sensitive information through visualizations, which provided valuable insights for data pre-processing and data manipulation in our project.

The paper *Crime Analysis and Prediction using Big Data* by Aarathi Srinivas Nadathur et al. Kattankulathur (2018) initially reviews and identifies features of crime incidents proposing a combinatorial incident description schema. The authors demonstrate the use of big data analytics tools for analyzing large, voluminous data sets. It inspired us to take advantage of various analytical tools for manipulating big data and visualization, such as Sqoop or Hive, and also motivated us to keep learning new techniques that can be applied to an industrial level in the future.

3. Data

3.1. About the Dataset

The dataset reflects incidents of crime (with the exception of murders) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is retrieved from the Chicago Police Department, who collects and also owns the data, and updates the dataset daily. The dataset was created in September 20, 2011, and last updated in May 11, 2023. The dataset can be found at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

In this dataset, there are 7.78 million rows and 22 columns, each row represents a reported

crime. Each crime has a unique identifier, date when the incident occurred. Other information in columns include description of location, indication of whether arrest was made, which community area the incident occurred, and crime classification outlined in the FBI's National Incident-Based Reporting System. As we can see, most of the content in the dataset is displayed in plain text, which brings about a big challenge in this project to deal with such data type and extract effective information from them.

3.2. Data Import

We used the API provided in the Chicago Data Portal to import the data. Considering the large volume of data size and to make our analysis more up-to-date, we only imported data starting from January 1, 2020. This also gives us a closer look at the impact of COVID-19 on crime types.

Until the last day updated, the imported dataset contains 732,743 observations of 22 columns. However, as we are connecting a dynamic source, the data size may come up to a limit of 1,000,000 one day with thousands of rows daily added to the data. To ensure our imported data starts from 2020, we take a step further to dynamically update the limit of data size. That is, if the imported data reaches the limit of 1,000,000, the procedure will automatically raise the limit by 500,000 for each time. This makes our design of data import more robust and steady.

The original data is in the JSON form, we convert it to a dataframe and stored it in CSV file named as "data", and the first few observations are displayed as below:

	id <chr>	case_number <chr>	date <chr>	block <chr>	iucr <chr>
1	12016034	JD193556	2020-01-01T00:00:00.000	018XX N WINNEBAGO AVE	1153
2	12220321	JD430436	2020-01-01T00:00:00.000	091XX S DREXEL AVE	1752
3	12013828	JD191019	2020-01-01T00:00:00.000	044XX S LAVERGNE AVE	0281
4	12019692	JD197444	2020-01-01T00:00:00.000	032XX N LINCOLN AVE	1153
5	12843813	JF415893	2020-01-01T00:00:00.000	022XX E 70TH ST	1153
6	12036792	JD216459	2020-01-01T00:00:00.000	072XX S WHIPPLE ST	1154

6 rows | 1-6 of 22 columns

Figure 1: Head Data

3.3. Data Pre-processing

As most of the variables are factors with characters, which contains duplicated records and unorganized categories, we made three majot changes to the data set at this stage.

First, we converted some columns to their respective types based on the properties of the variables in the dataset. We used `as.factor()` to convert the columns that indicate geographical information into categorical variables.

Second, we identified duplicate records that have the same `case_number` and removed them from the data set. 144 records were deleted through this step.

Lastly, we reorganized the crime types. The data contains 33 crime types; not all of which are mutually exclusive. We combined two or more similar categories into one to reduce this

number and make the analysis a bit more manageable. Specifically, we converted "CRIM SEXUAL ASSAULT", "PROSTITUTION", "SEX OFFENSE" to "SEX", and converted "NARCOTICS", "OTHER NARCOTIC VIOLATION" to "DRUG" and convert "PUBLIC INDECENCY", "RITUALISM", "HUMAN TRAFFICKING" to "OTHER" cases. Now we have cleaned our data.

4. Exploratory Data Analysis

In Exploratory Data Analysis (EDA) the first and foremost step in data analysis process. Here, we try to make sense of the data, figure out patterns, trends, outliers, try to form questions and as well as find out the best ways to manipulate the available data sources to get the answers needed.

4.1. Numerical Summaries

The cleaned dataset now contains 738595 rows and 24 variables, most of which are categorical ones indicating geographical information, like in which district or community area a crime occurred. Numerical variables include longitude and latitude of the location of a crime. From conclusion on frequency of crime types, it's explicit to see top ranking crime types as theft and battery. We are also interested in the categorical factor "arrest" indicating whether an arrest was made. The ratio of arrest under crimes is approximately 13%, and we are curious about the reason behind such a low percentage.

```
  arrest       domestic          beat          district                crime_type
FALSE:641280   FALSE:597451   1834   :  8143   006    : 47301   THEFT              :153530
TRUE : 97315   TRUE :141144   0421   :  6621   011    : 46679   BATTERY            :136150
                              0624   :  5765   008    : 45978   CRIMINAL DAMAGE    : 86302
                              0511   :  5675   004    : 43328   ASSAULT            : 66170
                              0123   :  5382   012    : 40276   DECEPTIVE PRACTICE : 56716
                              0631   :  5371   025    : 38273   MOTOR VEHICLE THEFT: 51582
                              (Other):701638   (Other):476760   (Other)            :188145
```

Figure 2: Numerical Summaries

4.2. Plot 1

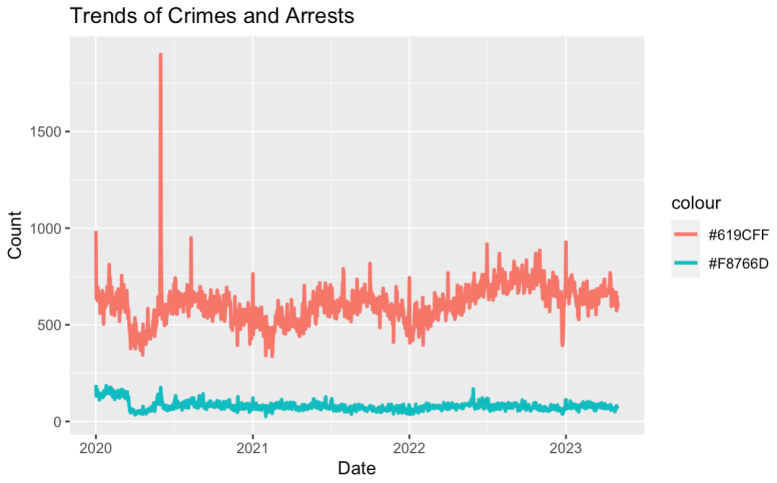


Figure 3: Trends – Crime – Arrests

This figure shows more crimes were committed but less arrests were made. The majority of daily reported crimes fell between the range of 500 to 1000, but there were a few outliers that exceeded 1800 incidents in a single day. Notably, there was a significant drop in crime reports during March and April of 2020, and again in January and February of 2021. The number of arrests per month remains more stable, never exceeding 250.

4.3. Plot 2

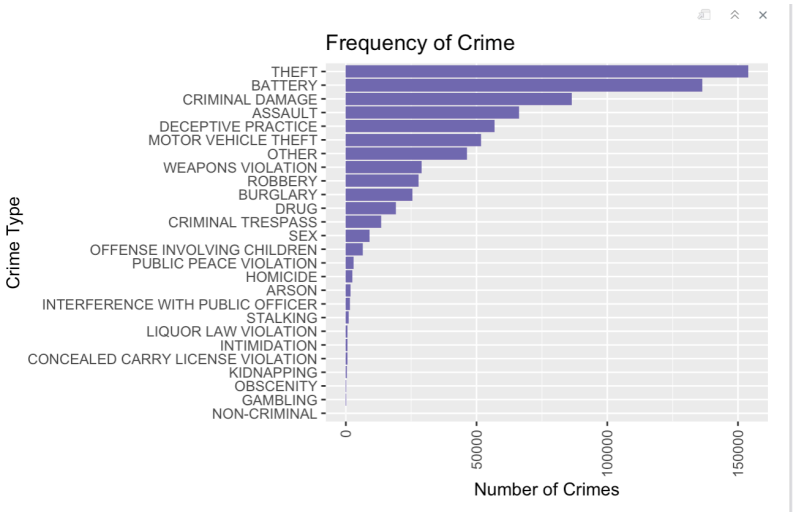


Figure 4: Crime Frequency by Criminal Type

Presented in this bar chart are the frequencies of different types of crimes, specifically those with reported cases exceeding 1,000 in 2020. The prevalent types of crimes were theft, assault, criminal damage, and battery. Among these, theft was the most frequently reported, with over 150,000 cases.

4.4. Plot 3

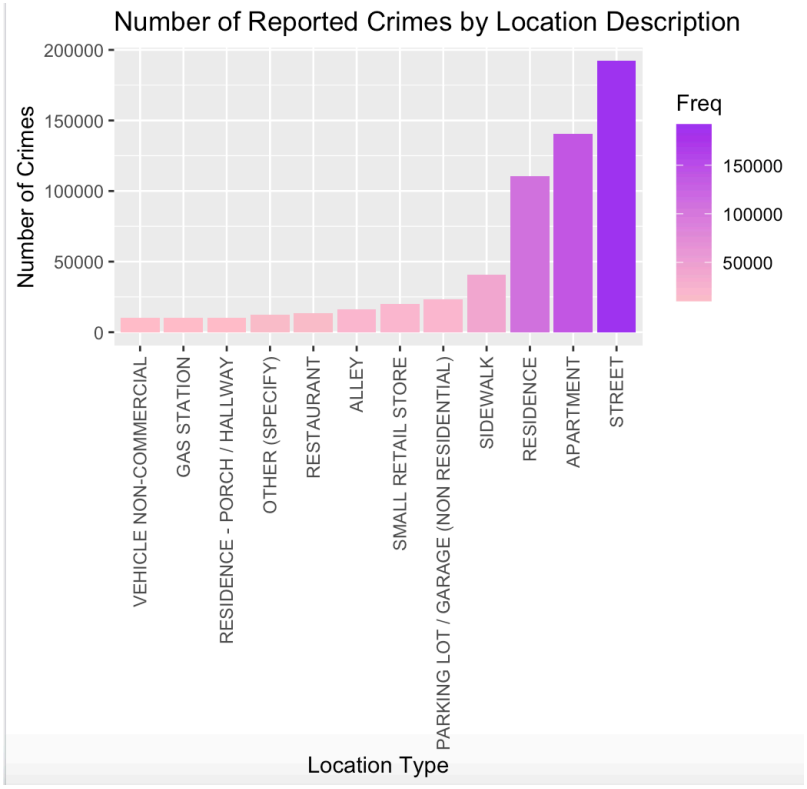


Figure 5: Crimes by Location Description

This bar chart shows number of crimes by criminal location, and we selected out locations with reported cases greater than 10,000 in here. Top ranking places include street, apartment, and residence.

4.5. Plot 4

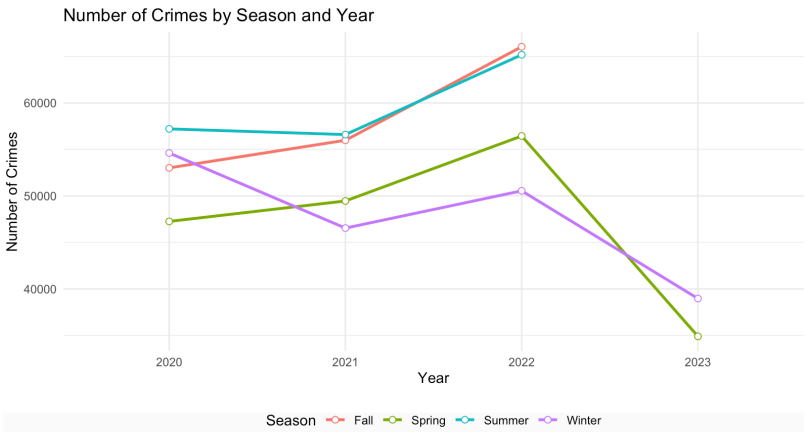
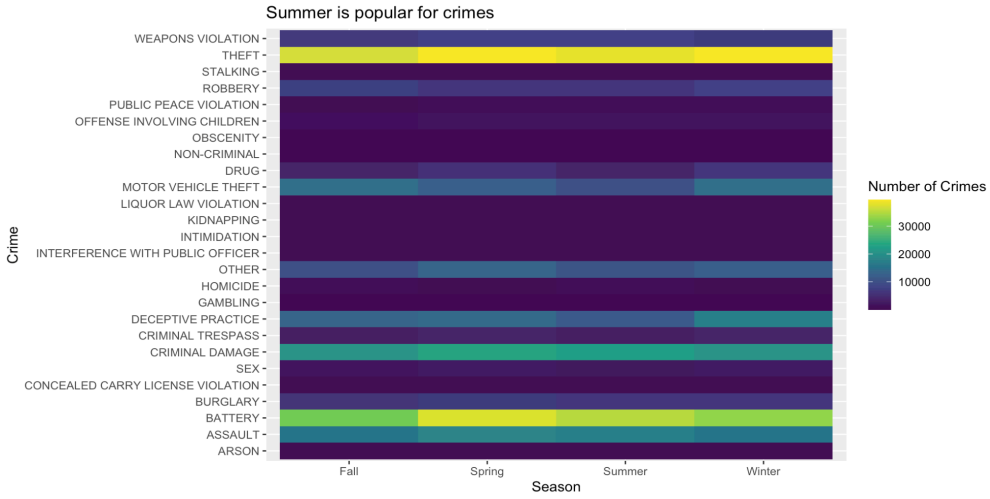


Figure 6: Crimes by Season and Year

We can see number of crimes is generally larger in fall and summer, and shrinks in winter and spring, which indicates seasonality within. Annually speaking, the reported cases in 2022 are comparably much more than that in 2021 and 2020.

4.6. Plot 5



The heat map indicates that theft is the predominant type of crime in Chicago across all seasons. Battery also occurs frequently throughout the year, with a peak in the spring. Although not as frequent as theft and battery, criminal damage is still a significant type of crime and is fairly uniformly distributed across all seasons.

4.7. Plot 6

Figure 7: Heatmap for Frequency by Crime and Location

This figure displays a Heat Map revealing that theft crimes are frequently committed in apartments and on the streets. Criminal damage and motor vehicle theft crimes are predominantly reported on the streets as well. In residential areas, battery, deceptive practice, and theft are frequently happened. In contrast, alleys are the locations where fewer crimes tend to occur.

Based on the visuals above, the data set suggests that crime is a significant problem in Chicago, with a complex set of factors contributing to the number and types of crimes reported. There appears to be a clear seasonality to the number of crimes reported, while number of arrests stay low. We are curious about the factors behind this and constructed a more detailed analysis.

5. Predictive Analysis

5.1. Model description and data pre-processing

We used logistic regression to predict whether an arrest was made or not under a certain crime. Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no.

Other classification models like KNN or random forest can also be applied to the prediction. Considering limited number of variables and most of them are categorical variables with multiple levels, we chose logistic regression to fit our model. We sample 80% of records from

the dataset to be training data, and the 20% left to be testing data.

5.2. Feature selection and evaluation

The explanatory variables chosen for prediction based on the pre-processing done during the exploratory analysis are as following:“location_description” “beat” “district” “ward” “community_area” “latitude” “longitude” “crime_type” “arrest” “season_summer” “season_winter” “season_spring” “season_fall”. Here we used One-hot encode the categorical variables into dummy variables and encoded categorical variables into numerical ones.

For evaluation, we first applied correlation matrix to these variables to check their correlation. For predictions, we set threshold equal to 0.5. For predicted probabilities larger than 0.5, we consider it as arrest made (1); for probability less than 0.5, we predict it to be 0. To further access the model’s performance by constructing a confusion matrix. Here’s how the confusion matrix works:True Positives (TP): It represents the number of instances correctly predicted as positive by the model. True Negatives (TN): It represents the number of instances correctly predicted as negative by the model. False Positives (FP): It represents the number of instances incorrectly predicted as positive by the model. Also known as Type I error. False Negatives (FN): It represents the number of instances incorrectly predicted as negative by the model. Also known as Type II error. We mainly focus on Accuracy: It measures the overall correctness of the model’s predictions, calculated as (TP + TN) / (TP + TN + FP + FN). It helps in understanding the types of errors made by the model and can aid in fine-tuning the model or adjusting decision thresholds to optimize its performance.

6. Results and Discussion

6.1. Predictive analysis

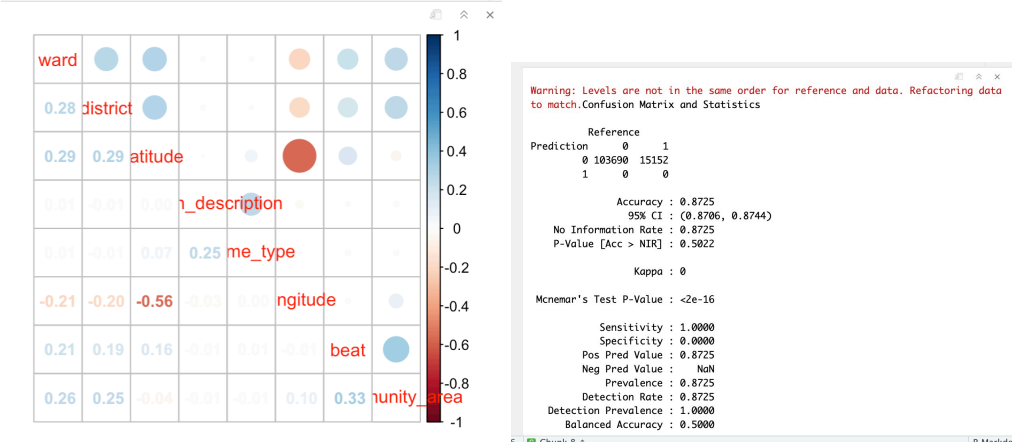


Figure 8: Evaluation Results

The correlation between explanatory variables are not significant so we keep all 12 variables to make predictions.

From the performance matrix, we can see the model reaches an accuracy of 0.8725, which is a decent score. However, taking a closer look at the predicted results, we found that the model is only making predicted values to be 0. This may because of our setting of threshold. As it can be directly seen from original data that the ratio of arrest/crime is about 13%, so the model accuracy is somehow interpretable and understandable. However, when we tune the threshold to be 0.13 or 0.3, the accuracy it achieves all drop below 0.87. Therefore, we keep threshold to be 0.5 in this case.

6.2. Shiny application

On top of that, we also built a dashboard for user interactions and show how predictions change when we change input variables.

The dashboard contain three major functions: 1. Show all 6 EDA graphs; 2. Summary logistic model statistics given different combination of variables where user can select and run the model himself; 3. Plot of fitted model.

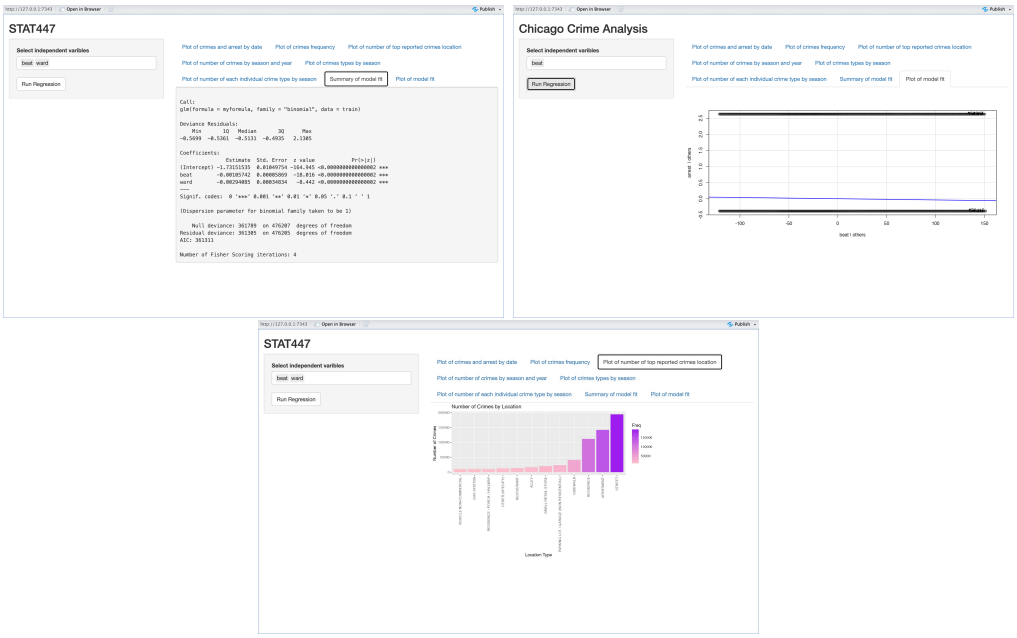


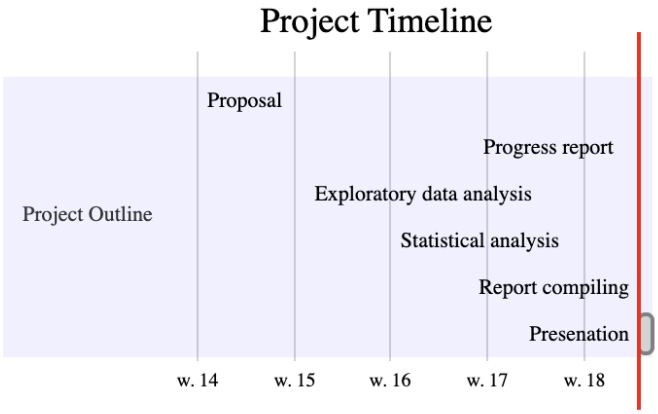
Figure 9: Shiny Application

7. Conclusion and Future Work

In EDA, we explored various criminal frequencies under different time, location or types, and the arrest rate under certain conditions is what we are curious about. To better understand the question, we built a logistic regression to predict arrest rate under different variables. We found that including all explanatory variables in the EDA process such as location description, beat, community area and season can give a decent prediction on arrest rate, which is close to true value.

However, there are also perspectives we haven't done yet. For example, the power of explanation of selected variables. We filtered 12 variables our of 24 columns, most of which indicate geographic information and locations, but there're still some different aspects we can interpret the model. For example, the correlation between arrest rate and domestic violance. On top of that, we can use more approaches to encode the categorical varibles to make it more random, and more statistical techniques like lasso or ridge for feature selection. These are probably the work we are going to do in the future.

8. Timeline



Yu was responsible for EDA, Predictive analysis, constructing shiny app and package. Yijia was responsible for compiling documents and presentation.

9. Contribution

Yu Wang is responsible for: data import and EDA, modelling, shiny, package, and correspond-

ing parts in writing report and general revisions. 70% Yijia Hu is responsible for: abstract, introduction and literature review, data storytelling, presenation. 30%

References

Kattankulathur K (2018). "Crime analysis and prediction using Big Data." *International Journal of Pure and Applied Mathematics*, **119**(12), 207–211.

Mallula R, Chowdary P (2018). "Predictive policing in crime analysis using R." *International Research Journal of Engineering and Technology*, **5**(7), 462–465.

Monish N (2019). "Chicago Crime Analysis using R Programming."

Affiliation:

Yu Wang (yw78)
Department of Statistics
E-mail: yw78@illinois.edu

Yijia Hu (yijiahu3)
Department of Statistics
E-mail: yijiahu3@illinois.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd