Ibraheem Saleh
CSE 482 (Spring ‘19)

The goal of this project is to develop a framework for crime prediction using diverse types of data available. Initially, the plan was to use data from different cities to provide a more diverse data pool. After further analysis, I have decided to focus on one data set specifically (Chicago) as it is already diverse and allows for a more specific analysis of data points. The source is monitored regarding future changes and those have been factored in. Crime data changes every day so allowing the project to analyze new incoming data will be a more accurate analysis in the end. The same timeline will be used for all the cities in the end to keep data analysis consistent, as well. Specifically; given a Date (exact date or year), Block, Domestic, Primary Type, Location, and Location Description; a prediction of the likelihood that a crime will occur can be obtained. If the model developed is accurate it can then be applied to all locations in a region for a future time period.

The data elements being used will be Date, Block (area of occurence), Domestic (True/False), Primary Type (type of crime based off set labels such as SEX OFFENSE, ROBBERY, DECEPTIVE PRACTICE, CRIMINAL DAMAGE, OTHER OFFENSE), Location (district number and ward number), and Location Description (house, apartment, street, etc). With the large amount of data collected by the Chicago Police Department, detailed analysis can be done on the varying neighborhoods throughout Chicago. Currently, I have sifted through the data sample provided by the Chicago government from the year 2016 to the present date and organized the data based off the previously listed data points (pre-processing). After successful organization of the data, I will target a larger date range (2010 to 2019) and eventually address the entire data set provided (2001 to 2019).

The main technique planned on being used is testing sets from the regression models to predict future crime occurrences. Data will be divided to test for separate crime types and will be used to create figures to analyze correlations between different crime type occurrences and timing, as well as location. A large problem with this specific project is missing data and the possibility of certain crimes being wrongly reported.

The type of data sets used in the current project are based off the data elements being used: Date, Block, Domestic, Primary Type, Location, and Location Description.

Ibraheem Saleh
CSE 482 (Spring '19)

ID	Case Number	Date	Block	Primary Type	Location Description	Domestic	District	Ward
10606686	HZ358366	6/20/16 9:00	047XX S INDIANA AVE	SEX OFFENSE	RESIDENCE	FALSE	2	3
10750326	HZ512286	11/12/16 7:30	038XX W 61ST ST	ROBBERY	SIDEWALK	FALSE	8	13
11094370	JA440032	9/21/17 0:15	072XX N CALIFORNIA AVE	DECEPTIVE PRACTICE	CURRENCY EXCHANGE	FALSE	24	50
11118031	JA470589	10/12/17 19:14	055XX W GRAND AVE	CRIMINAL DAMAGE	JAIL / LOCK-UP FACILITY	FALSE	25	29
11134189	JA491697	10/30/17 11:52	043XX S TALMAN AVE	OTHER OFFENSE	APARTMENT	FALSE	9	12
11156462	JA521389	9/29/17 18:45	055XX W BELMONT AVE	DECEPTIVE PRACTICE	CURRENCY EXCHANGE	FALSE	25	30
11164874	JA531910	12/1/17 6:20	022XX W CHICAGO AVE	CRIM SEXUAL ASSAULT	STREET	FALSE	12	32
10974196	JA298243	6/8/17 16:02	001XX E 59TH ST	WEAPONS VIOLATION	ALLEY	FALSE	2	20
10995008	JA322389	6/25/17 19:00	003XX W 103RD ST	ROBBERY	SMALL RETAIL STORE	FALSE	5	9
11086755	JA430240	9/6/17 13:30	032XX W 26TH ST	THEFT	OTHER	FALSE	10	12
11162428	JA529032	11/28/17 21:43	026XX S CALIFORNIA BLVD	OTHER OFFENSE	JAIL / LOCK-UP FACILITY	FALSE	10	12
11175304	JA545986	12/11/17 19:15	007XX N SACRAMENTO BLVD	ROBBERY	SIDEWALK	FALSE	12	27
11175934	JA546734	12/12/17 18:00	007XX N OAKLEY BLVD	ROBBERY	STREET	FALSE	12	32
11230118	JB151212	11/22/17 21:08	003XX W 79TH ST	DECEPTIVE PRACTICE	CURRENCY EXCHANGE	FALSE	6	17
11172603	JA541446	12/8/17 3:00	033XX S KEDZIE AVE	ARSON	SCHOOL, PRIVATE, BUILDING	FALSE	10	22
11223717	JB141803	2/4/18 15:25	044XX W WASHINGTON BLVD	DECEPTIVE PRACTICE		FALSE	11	28
11227287	JB147188	10/8/17 3:00	092XX S RACINE AVE	CRIM SEXUAL ASSAULT	RESIDENCE	FALSE	22	21
11227293	JB147230	9/9/17 20:17	060XX S EBERHART AVE	THEFT	RESIDENCE	FALSE	3	20
11227475	JB147314	11/22/17 2:42	056XX N CHRISTIANA AVE	OTHER OFFENSE	APARTMENT	TRUE	17	39
11227505	JB146342	6/1/17 0:01	115XX S ABERDEEN ST	OFFENSE INVOLVING CHILDREN	RESIDENCE	FALSE	5	34

The above image is the beginning of a very lengthy table I am using with the variables I will not be using from the initial csv removed. Only the essential data remains in my table and is being used to process and will (hopefully) calculate a promising prediction of if/when a crime will occur and possibly what type of crime is most likely to be occurring at that time.

Timeline:

- Feb 11 - 15
  - Setup timeline and organize data collection system
- Feb 18 - 22
  - Continue organizing data and collecting consistent findings
- Feb 25 - March 1
  - Specify crime types to track in Chicago to keep data consistent
- Mar 4 - 8 (Spring Break)
  - Find relationships between crime data types
- Mar 11 - 15
  - Begin processing data based on each primary type and compare with larger data sets
- Mar 18 - 22
  - Intermediate Report due March 24
  - Begin processing correlations between different crime type data analysis
- Mar 25 - 29
  - Divide data into specific sections for easier organization and display flow
- Apr 1 - 5
  - Evaluate results from data analysis and fix any errors and inconsistent findings
- Apr 8 - 12
  - Continue any required debugging and error fixes
- Apr 15 - 19
  - Finalize prediction model
  - Final report due on April 21 (finish final report and findings)