California State University Northridge

Los Angeles and Chicago Crime Data Analysis

Group 4

Sean Hansen

Alonso Gonzalez

Jose De Ita

Spring 2018

05/10/2018

COMP 541 Data Mining

Course ID: 16399

[**Objective**](#_rsamj0smjjgm) **2**

[**Background Information**](#_j0d1b2j7wjzc) **3**

[**Design**](#_w963xgam4o89) **5**

[**Implementation & Results**](#_dxu5xeiqiopo) **6**

[**Result Analysis**](#_ho020uysnbuh) **9**

[**Retrospective**](#_ccz7qfbuh0d3) **9**

[**References**](#_vl80ifu5owlm) **10**

[**Appendix A**](#_d9pn32jid518) **1**2

# 

# Objective

The objective of this project is to analyze crime data from the cities of Los Angeles and Chicago, which cover the years 2010 - present and 2001 - present, respectively. In doing so we will practice several data mining techniques including but not limited to: cleaning/preprocessing, data warehousing, OLAP techniques, and various mining strategies.

We will attempt to clean and preprocess the data, including filling in missing values as well as performing reduction techniques. Since we believe that the only viable numerosity reduction that suits our data is aggregation, we will also create a data warehouse to store our data with which we will create a data cube.

We will also attempt to establish and visualize trends in both datasets, not only in overall crime but on an individual type basis. By mining frequently committed crimes, we will attempt to establish association rules between crimes to give an idea of what crimes are usually committed together and with what support/certainty. We also plan to view crime trends from the multiple dimensions available to us including: time, premise, location, and for the Los Angeles dataset, victim information.

Finally, we will implement various mining techniques described in class including decision tree prediction, naive Bayes classifiers, and neural networks. With these we hope to view how crime is determined from its multiple dimensions. With luck we will also be able to create reliable models with which we can predict crime types, future trends, and various other attributes from our data.

# Background Information

We will extract our data on crime in Los Angeles and Chicago from their respective city portals. These portals provide public access to these cities’ extensive crime databases, which contain crime data dating back to 2001 for Chicago, and 2010 for Los Angeles. Each portal allows the public to download the datasets in CSV format. Links to these portals can be found in the appendices to this report.

A brief description of each dataset in its raw, unaltered from follows:

|  |  |
| --- | --- |
| **Los Angeles** | **Chicago** |
| * Rows: ~6.5 Million * Attributes: 21 * Example Attributes:   + ID (Primary Key) - Integer   + Case Number - String   + Primary Crime Type - String   + Crime Description - String   + Location Description - String   + Arrest Made - Boolean   + Domestic Crime - Boolean   + Geographic Coordinates - Set of Real | * Rows: ~1.6 Million * Attributes: 26 * Example Attributes:   + Date Reported - Datetime   + Time Occured - String   + Crime Description - String   + Victim Age/Sex/Descent - Int/Ordinal   + Location Description - String   + Address - String   + Geographic Coordinates - Set of Real |

While the format of each dataset differs in several aspects, in general each one follows a similar format. Each row represents a distinct crime that occurred, and each column represents specific data about that particular crime. Each dataset, at minimum, contains the following information:

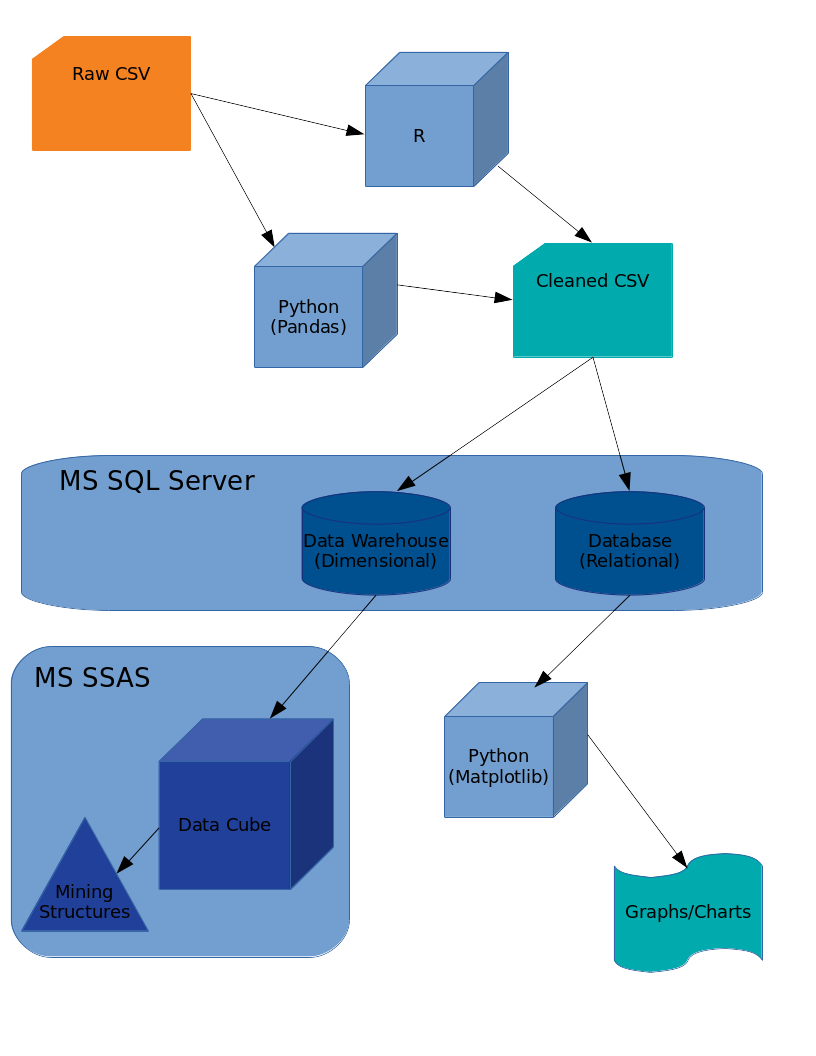
* A unique identifier
* The time that the crime occurred (to the minute)
* The geographic area where the crime occurred
* A city specific crime code and description
* A premise description of where the crime occurred

However this is where the similarities end, as each city defines its own crime codes and descriptions, its own premise descriptions, and each provides its own additional attributes. For example and of particular interest to us, Los Angeles also provides victim information for each crime including their sex, descent, and age.

To perform our analysis we will use several tools listed below:

* Cleaning / Preprocessing
  + **Pandas** - An extremely powerful and robust data manipulation library for Python, which we will use to perform the majority of the project’s cleaning/preprocessing tasks
  + **R** - A very popular language among data scientists, will be used for some cleaning/preprocessing tasks.
* Data Storage
  + **Microsoft SQL Server** - Our main database used to store cleaned data, as well as our data warehouses. We used an AWS RDS instance to store some of our data as well as our individual machines.
* Analysis Server / OLAP
  + **Microsoft SQL Server Analysis Services (SSAS)** - Microsoft’s very popular tool for data mining/business intelligence. We chose SSAS as it integrates well with MSSQL server.
* Visualization / Documentation
  + **Jupyter Notebook** - A popular tool for sharing python code and output/visualizations. We will use this to show our techniques for cleaning/preprocessing as well as visualizations.
  + **Matplotlib** - A useful graphing library for Python, which we use for the majority of our data visualizations.

# Design



# Implementation & Results

Missing Data: Los Angeles

The number of missing attributes in the dimensions we were interested in had to be dealt with before any further analysis could continue. Missing descriptions were recovered by mapping the codes into their representations. The real issue was how to deal with the near 10% missing data in the victim information dimensions.

Victim Age, Sex, Descent were missing in 150,000 records of the 1.7 Million. We wanted to add these values without distorting the analysis results. Since age could be mined in ranges, we felt that a simple mean average based on grouping record by crime and premise was suffice to get rid of all nulls in that field. In contrast, Sex and Descent only have mode to consider and we tried running similar preprocessing methods but the data is biased towards male hispanics in parts because of geographical reasons. Ultimately, we decided to place an X in these fields, the X stands for unknown to which definition was inherited from the original data set. (more in Appendix A)

Data Warehousing: Los Angeles

Originally, we had stored all of our data on a sql instance using Amazon’s Warehouse Service. This brought us a uniform view of the data for all of us. A big problem was that we didn’t have enough permissions to perform anything other than read and write operations on the database. Midway through the semester we scrapped using AWS in favor of Microsoft’s Sql Server. With local instances of sql running we were able to use SQL Server Data Tools(SSDT) for building data cubes, OLAP operations, market basket analysis, applying a Decision Tree Algorithm, and, Bayesian Classifiers.

We were unable to make a data cube for the Los Angeles data.

Types of data sets for Chicago:

Chicago’s data set consisted of mostly nominal data. There was one date attribute and a few discrete numerical data. The discrete numerical data consisted of locations information, such as coordinates, and year. There were also two attributes that were binary nominal data. They were arrest and domestic. With a 1 representing an actual arrest or a domestic violence and a 0 representing no arrest or no domestic violence. We also had two continuous numerical data attributes. Those being latitude and longitude.

Missing Data: Chicago

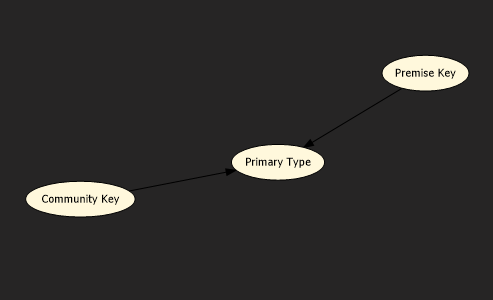
Chicago’s missing data was surprisingly uniform. There were only 10 attributes with missing data. Usually, the missing data came in groups. For example, x coordinate, y coordinate, latitude, longitude, and location all had 96017 missing values. This wasn’t such a big problem since those fields are derived from other fields. We decided to drop those fields in favor of Police area, ward area, and community area. These fields would give us an understanding of how each separate community in Chicago was being afflicted by crime.

Data Warehousing: Chicago

Like we mentioned earlier we used sql server as a data warehouse. The data was stored using a star schema. Our fact table consisted of a primary key ID, Date, IUCR(crime codes for chicago), Primary Type(general crime type), Description(description of the crime), Premise Key, Arrest, Domestic, Police Key, Ward Key, and Community Key. This table consist of 6,440,380 tuples and takes up 715.461 megabytes of space. The other tables were a time table, police table, community area table, crime type table, ward area table, and a premise table. Time table is a simple table consisting of time tuples, which help with mining. Police table has the different police districts from Chicago. The community area table contains the different community areas of Chicago.

Data Cube: Chicago

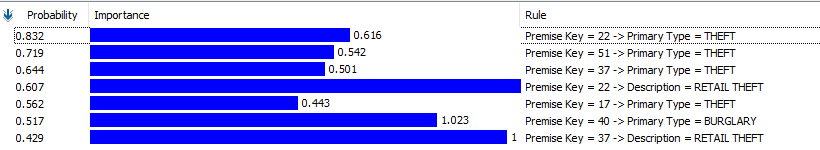
We made the Chicago data cube using SSDT. We used Time, Premise and Crime type as the dimensions of the data cube. The reason for those dimensions are that we wanted to view the crimes from a time perspective, a location perspective and the type of crime committed. With the cube built, we performed drill down operations from all to year to quarter to month. This gave us an idea of what type of crimes were being committed at different times.

Decision Tree Algorithm: Chicago

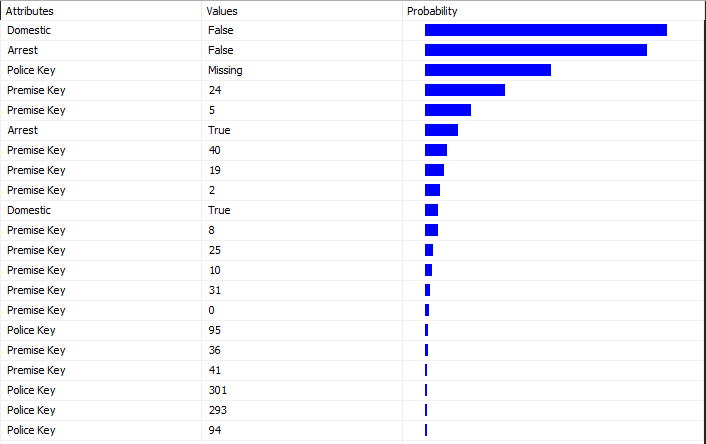
Our decision tree for Chicago was very limited. This was due to the fact that most of our data is categorical. The curse of high dimensionality became very apparent with this method as well. Even with a simple mining structure, made up of Community key, primary type and premise key, we had a tree with 44 tree levels. In fact, the tree was so large that the tree couldn’t be viewed completely. Another issue with our tree was that it took 15 minutes to build.

Association Rules: Chicago

We had a few association rules with high enough probably to be considered interesting. As you can see below, all of the association rules we found had to do with theft of some kind. The first three rules have premises department store, drug store, and grocery store and the associated primary crime type is theft. There is also a rule for “RESIDENCE - GARAGE” and burglary, with a probability of 51.7%.



Bayesian Classifier: Chicago

Bayesian classification for the Chicago data was also performed using SSAS on our previously constructed data cube, using Microsoft’s standard naive Bayes classifier. We chose to only use a small subset of the attributes, and decided upon community area and premise to be the features of our classifier. We chose primary crime type as the prediction target of our classifier, and we trained the model on 30% of the data.

# 

# Result Analysis

Chicago Analysis:

The key important thing about our analysis is that, overall, crime is going down in the city of Chicago. There were almost half a million crimes reported in 2001. By 2017, the number was slightly above 250,000. Homicide is an interesting outlier crime in Chicago, as it sharply increased during the years of 2016 and 2017. It has since decreased, however it is still one of the only violent crimes that saw increases in occurrence during the past 5 years. The death toll reached 800 citizens for 2017. With the deaths mostly occurring in community area 25. The three highest reported crimes are theft, battery, and criminal damages. Criminal damages are essential damage to someone’s property without their consent.

Los Angeles Analysis:

The analysis of the Los Angeles data show that crime had decreased to very low levels between 2014 and 2015 but has dramatically increase in the last few years, at least the reporting has. It is not clear why crime reporting dipped in 2014, further analysis needs to be conducted. The most frequent crimes reported where battery, and burglary of and from vehicles.

# Retrospective

Police reports are prone to data manipulation, whether it be to protect the victims or criminals, we felt this type of data is too impure and leads to no interesting results to be mined. The only thing that was interesting to some extent was information on the frequency and types of crimes in these cities considered and it would be interesting to use this with other types of data using time as the common factor.

# References

1. EDA of Crime in Chicago (2005 - 2016) | Kaggle. (n.d.). Retrieved from <https://www.kaggle.com/fahd09/eda-of-crime-in-chicago-2005-2016>
2. Chicago Crime Database <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data>
3. Los Angeles Crime Database

<https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq/data>

# 

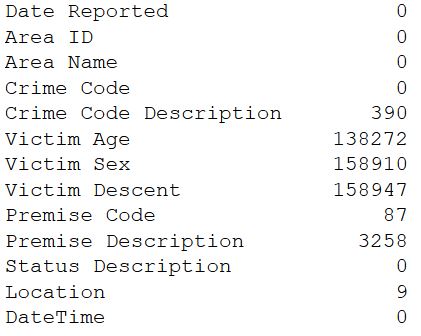
# 

# 

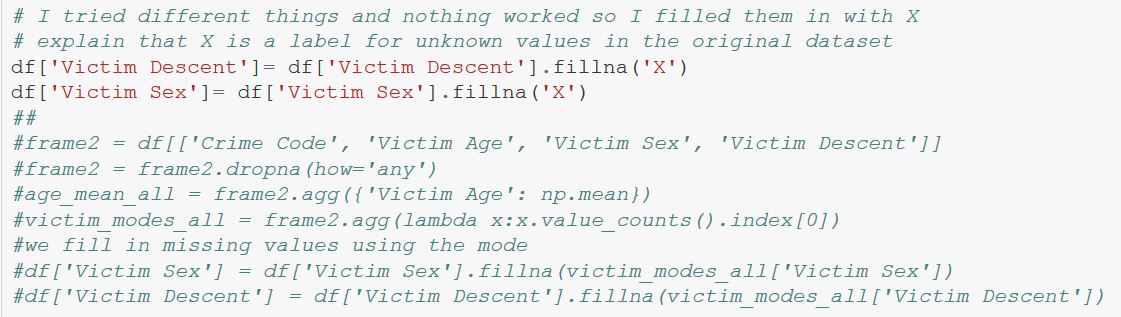
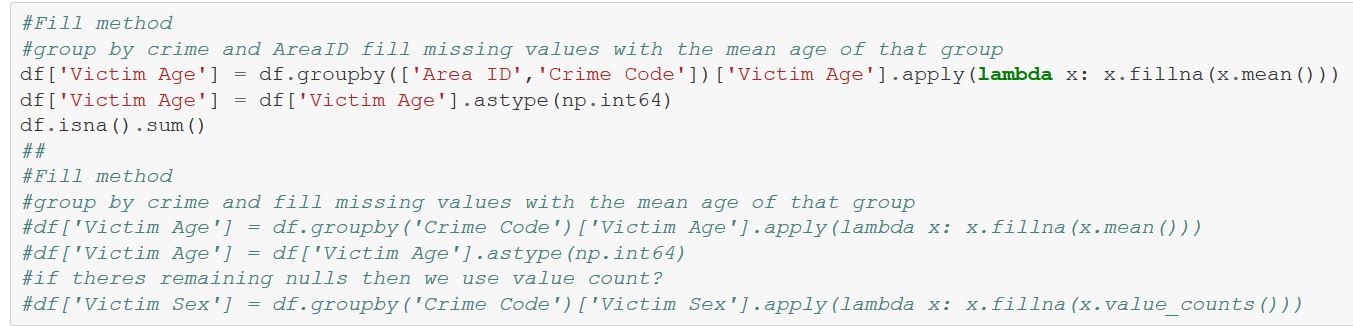
# 

# 

# Appendix

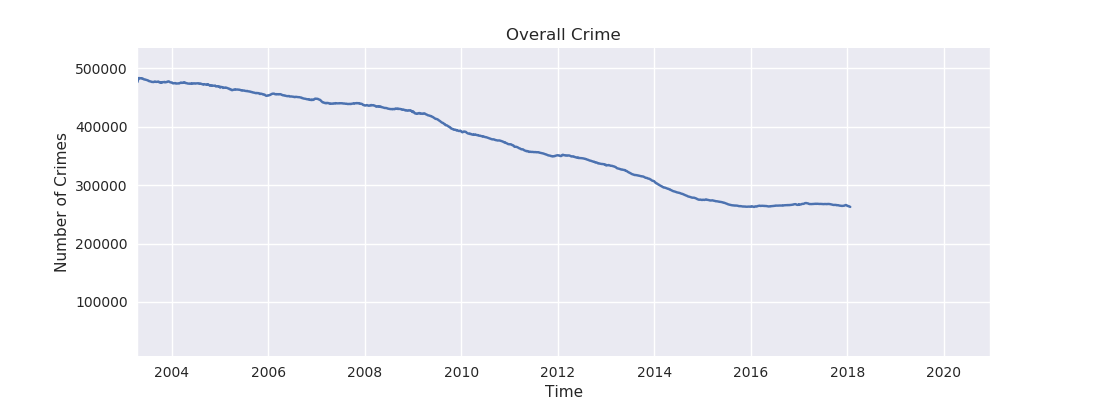
Missing data per column:

Python code for victim missing data:

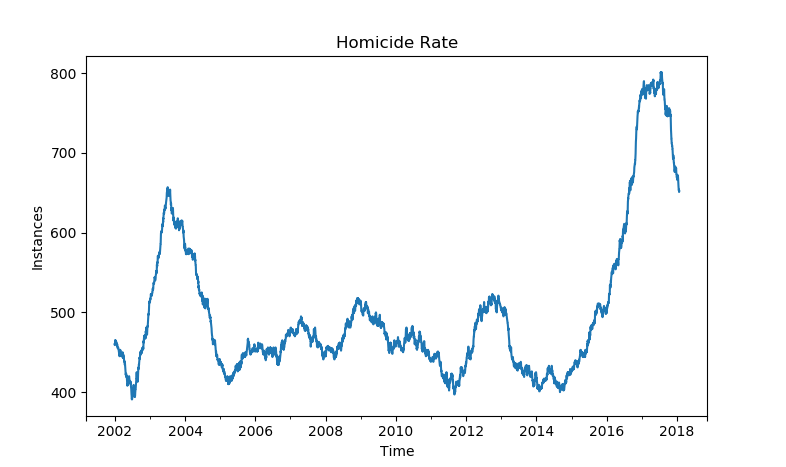


Python code for Descent/Sex missing data:

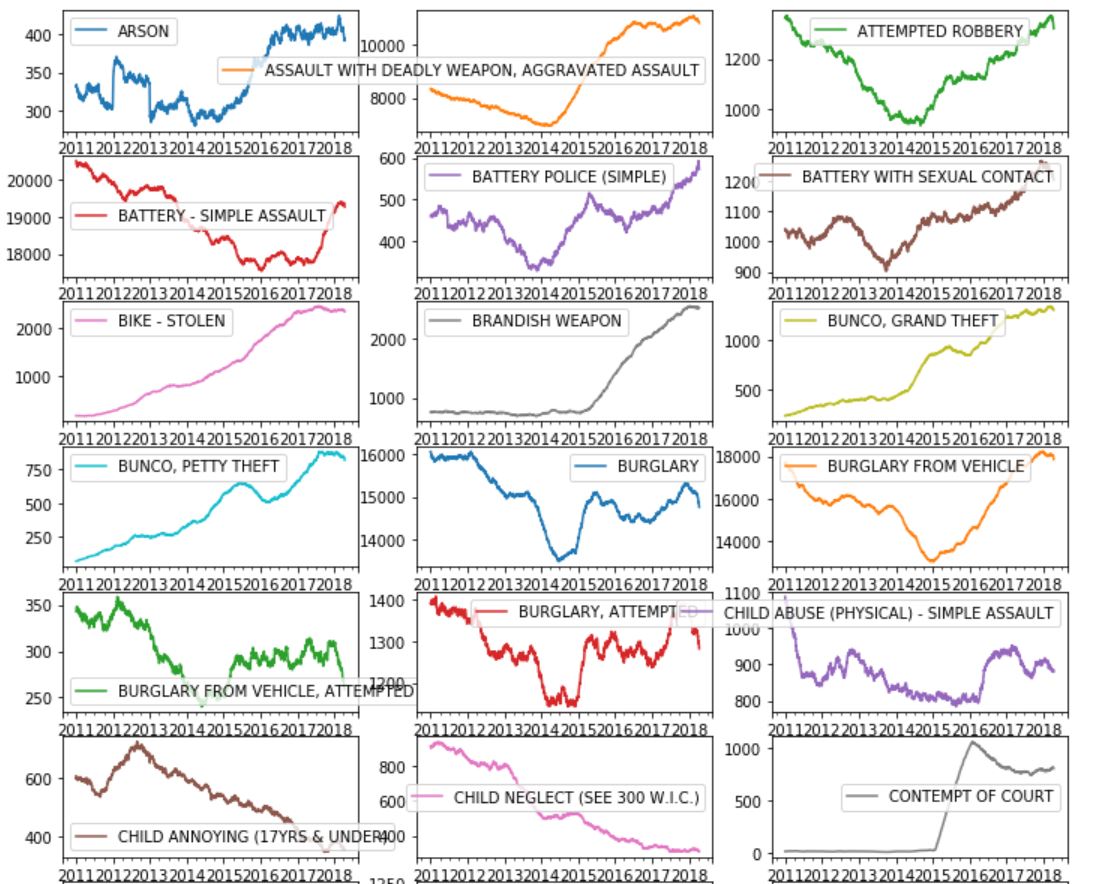
Overall crime trends in Chicago



Homicide rates in Chicago



CRIME RATES LA



LA CRIME RATES