Checkpoint 2: Data Integration
MSAI 339
Prof. Jennie Rogers
October 24, 2018

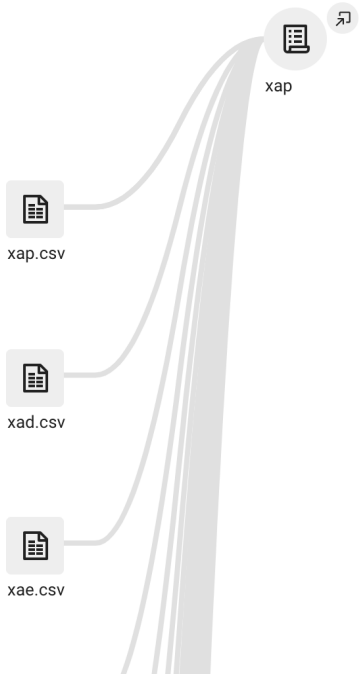
GROUP 5
Jack R
Quincia H
Ikhlas A

Purpose

The purpose of this checkpoint is to catalog the process of cleaning and integrating the Chicago Crime data with the Chicago Allegation data. This will be carried out through Trifacta Wrangler, a data tool that effectively cleans data sets for integration and reduce the complexity of queries.

Crime Data Cleaning

There is one large challenge with interacting with the crime data set. This challenge is that it is a single table with over seven million entries which results in downloading a 1.6GB CSV file for importing into Trifacta for cleaning. To circumvent the 100MB maximum import limit, the only viable option was to split the large table into seventeen smaller files with the same header columns. After using Mac's command line file splitting functionality, all seventeen data sets were loaded into Trifacta and put back together using a large union command. To clean the large data set, we removed columns that did not provide applicable information to the allegation data set.



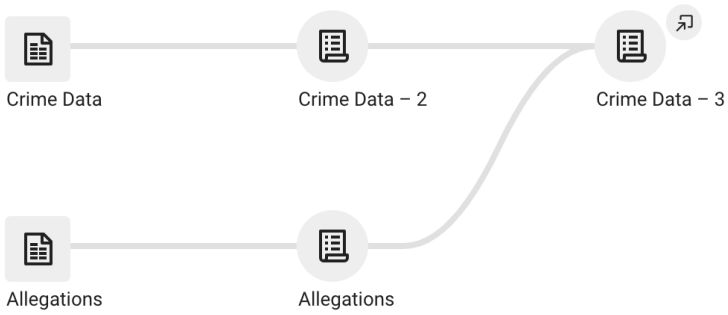
Allegation Data Cleaning

The priority of this checkpoint is to confirm that there is a viable connection between crimes and complaints. The first step in this is to isolate allegations into a table with only the nec-

essary columns. The necessary columns are AllegationID, IncidentDate, AddressNo, Address, and BeatID. AllegationID allows us to gain more information if there is a distinct connection. Incident Date, AddressNo, and Address are going to be used to connect to the Crime data. If there are allegations made on the same street and on the same date, there is a strong likelyhood there is a connection. This is especially true for days with very little crimes in a given area. To clean the data we imported the table to Trifacta, deleted unnecessary columns and split the timestamp to only worry about date. This is because the turn-around time of a crime to a complaint is unknown.

Data Integration

Below are the questions stated in the project proposal that will be discussed. In the event that a question could not be satisfied, or the results showed no correlation, a potential followup question will replace it to carry out before the next checkpoint.



Are there a significant number of complaints with near-equivalent entries of date and location in the crime database?

After cleaning up data in Trifacta, we were able to connect 118,932 allegations and crimes by date and address. This resulted in a join of 82% compatibility, ultimately removing any that did not combine. This was much better than expected, and will be beneficial in developing further analysis when we apply it further than allegationID to crimeID

column2	column3	column4	column5	column6
01/24/2008	2008-01-24	W 63RD ST	W 63RD ST	6038616
01/24/2008	2008-01-24	W GRAND AVE	W GRAND AVE	6038641
01/23/2008	2008-01-23	S HALSTED ST	S HALSTED ST	6038642
01/22/2008	2008-01-22	S MICHIGAN AVE	S MICHIGAN AVE	6038659
01/01/2008	2008-01-01	S THROOP ST	S THROOP ST	6038668
01/23/2008	2008-01-23	W 77TH ST	W 77TH ST	6038672
01/20/2008	2008-01-20	S HALSTED ST	S HALSTED ST	6038689
01/23/2008	2008-01-23	S PULASKI RD	S PULASKI RD	6038695
06/15/2007	2007-06-15	S MICHIGAN AVE	S MICHIGAN AVE	6038709
01/24/2008	2008-01-24	S STATE ST	S STATE ST	6038750
01/24/2008	2008-01-24	S STATE ST	S STATE ST	6038824
01/24/2008	2008-01-24	W GRAND AVE	W GRAND AVE	6038886
01/24/2008	2008-01-24	S STATE ST	S STATE ST	6039037
01/24/2008	2008-01-24	S STATE ST	S STATE ST	6039044
01/22/2008	2008-01-22	S HALSTED ST	S HALSTED ST	6039045

Given connections between crimes and complaints in a location at a given time, what classifications of crimes lead to a greater number of filed complaints?

This question was indirectly answered after answering the first question. Since Trifacta Wrangler shows distributions, we were able to identify the 3 largest crimes complained about: Theft,

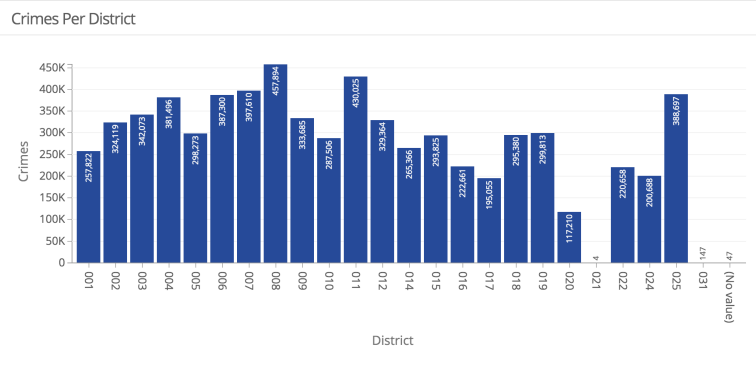
Battery, Narcotics. Our thoughts on this is that it somewhat makes sense due to the potential of denying any of those crimes in comparison to something like homicide.

Is there a stronger likelihood that a complaint will be filed in areas that historically show more domestic crimes?

This questions had disappointing results due the crime data set being heavily non domestic. More than 80% of the connected crimes were not domestic. Our analysis of this is that it stems from a large portion of the crimes were theft and narcotics.

What is the ratio of complaints to crimes per police district?

While this question served as a great direction for us initially, it does not provide anything beneficial other than the results received from checking the number of crimes per district. Our results for crimes per district are below. One thing we noticed is that there numerous crimes said to be in the 31st district, which does not exist.



Results and Analysis

Our results, while not as fruitful as hoped, did show a great connection between the two datasets. By joining on both Address and Date at the same time, we were able to pair 82 percent of the allegations with corresponding crimes. That being said, the number of crimes severely outnumbered the number of allegations, roughly seven million to 120,000 after factoring out all allegations before 2001. Now that we have connected CrimeID to AllegationID, we have plenty data to start working with in the future checkpoints.