CrimeIntel

A data-mining approach to public safety

Max Smith
maxs@colorado.edu

Aaron Shyuu
aaron.shyuu@colorado.edu

Thomas Dolan
thomas.j.dolan@colorado.edu

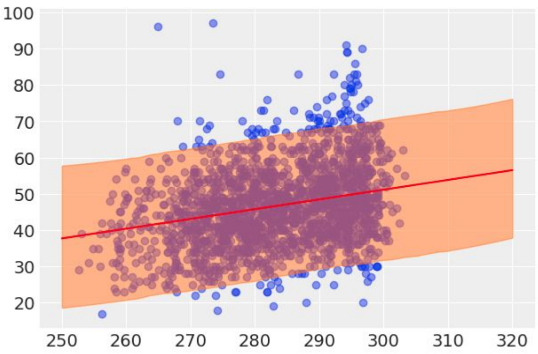
PROBLEM STATEMENT

Chicago has received national attention for its high homicide rates, especially since the rates skyrocketed in 2016 [1]. Not only has this attention precipitated a lot of interest in generating a solution, it has also shown that the causality of the problem has many roots, indicating that the impact of many proposed solutions have intractable results that are difficult to predict [2]. This has been further complicated by those purporting to have a grand solution, but whose primary goal is to support their own agenda. In order to provide possible solutions, we have to understand the nature of crime occurrences in Chicago with as much nuance, depth, and objectivity as possible. Using an objective, data driven approach is a likely means to that end. The size of Chicago and its crime data, collected by the Chicago City Police, lends itself to this approach. However, still missing is the very important community-driven data, such as surveys after an incident [2]. This means that any study on crime in Chicago will be inherently incomplete. For that reason, our goal is not to solve the problem, but to lend a new perspective that adds texture to the landscape that is our understanding of Chicago crime.

By using a unique combination of supplemental data-sets and other auxiliary information to enhance our data mining process, we hope to contribute to the conversation by answering smaller, more manageable questions and proffering any new, unexpected insights we might find along the way. Specifically, we hope to discover the correlation between crime and local economic trends. We will also be discovering whether homicide is correlated or associated with other types of crime. Breaking down crime types even further, we will inspect broader trends, such as whether property crime or violent crime have similar or different patterns from other crime-type categorizations. We will analyze the associations between locations and crime types to identify crime hotspots on a per crime basis, as a crime-type aggregate, and in general. Next, we hope to be able to delve deeper into crime type associations with more specific economic metrics. For example, we could explore whether there is an association between violent crime and unemployment metrics. Finally, using classification methods, we will be able to predict based on crime, time, and place, whether an arrest will be made, leading us to additional insights and future data mining questions.

LITERARY SURVEY

Many analyses relating to Chicago crime have been done including those that have used our same dataset. Some of these projects exist as notebooks accessible on Kaggle's website, while some others are hosted on personal blogs or sites. Four of these in particular provided inspiration for our project, two of which utilize data mining / KDD methods. The project "Chicago Crime on Christmas" helped us realize the benefits of integrating additional data types to yield novel insights, and the frequency of the data we would need in order to do so successfully [3]. A Python project by Cjango provided a vis heavy approach that helped us get a better sense for the data we would be working with and what kind of questions we could be asking [4]. Two separate data mining projects using different tooling than we plan to use were especially useful for understanding how we can apply the data mining methodologies we have been learning in class to our dataset [5, 6]. In both cases we found that they used only the categories provided originally in the data set and realized that if we were able to derive more attributes using additional categorization information, we could create hierarchies within the data that will help us yield additional insights.



Excerpt from Mikko Karkkainen's "Chicago Crime on Christmas" [3], showing a correlation between temperature and incidents of crime.

In addition to these projects, many academic and civic studies have also been done. A data driven article published by Northwestern helped give us context surrounding the problem, potential causes, and possible solutions [1]. An academic article by Julia Burdick-Will on the relation between violent crime and academic achievement in Chicago schools helped to illuminate the challenges associated with identifying causality and prescribing solutions to crime given its endemic nature and the role that data plays in this.

PROPOSED WORK

After the data collection process, we decided on the Chicago Crime dataset with the CFNAI dataset as a supplement. Before performing any Exploratory Data Analysis or data mining techniques, we need to ensure that our raw data are transformed into useful and efficient formats. Our focus for data preprocessing will be on the Chicago Crime Dataset, but the same procedures will apply to the CFNAI dataset. The Chicago Crime dataset is a huge dataset consisting of millions of rows. This means that the first step to cleaning our data will be to verify any missing/null values and address them appropriately. Through some initial filtering, it is clear that the majority of our missing values occur in the location and coordinate attributes. Due to the size of our dataset and the relatively small number of rows that contain at least a column with a missing value (< 0.1% of our dataset), we can safely remove them with minimal loss. Next, we need to check if there is any redundancy in our data and remove duplicate rows as they take up unnecessary space. We will be removing exact timestamps from dates since they overlap with other time data and we don't expect to inspect data at that level of granularity. To validate the quality of our data, we will review certain attributes that could be more prone to input errors, such as the fbi attributes, and make corrections if needed.

In terms of data reduction, we will remove attributes that are not beneficial to EDA or data mining such as case number and update time. Moreover, any ID attribute columns will be dropped as they do not contribute to the data mining process. We will iteratively remove attributes to narrow down our data further until all remaining attributes are essential. Since location attributes such as beat, district and ward are similar and might yield the same correlation, we will perform a Chi-Square test (correlation analysis) to see if we can reduce from those attributes. We will use four-year periods of the data for initial implementation of our pattern and data analysis to improve efficiency and faster runtime before applying those methods to the full dataset. Given the high number of categorical attributes and the fact that some of the more salient crime times will be underrepresented in the data, this is preferable to permanently reducing the data size with sampling or other methods that would be difficult to do correctly in this context without potentially affecting the results. As well, the efficiencies we hope to gain using pandas dataframes should enable us to finally apply our methods to the full dataset without issue.

To find correlations between the Chicago Crime dataset and economic dataset, it is crucial for us to integrate them. We will join the two datasets through the date attribute. This will give us a combined table where we will be able to see the values of crime attributes and economic attributes for a given date. (represented in an object or row of the table). We may also perform other integrations and aggregations as necessary to compare crime data against economic trends. In order to find correlations and associations, we will also make some transformations which includes rearranging our data into chronological order and grouping by crime types depending on the mining knowledge we are trying to gain. We will also derive nominal attributes—crimes against person, crimes against property, and crimes against society— which will be useful for multilevel pattern analyses and their relation to economic trends. Furthermore, new attributes such as fbi code descriptions will be created to support classification and model construction. For tactics involving a join on the CFNAI dataset, we will be performing an inner join on the full crime dataset, then perform reduction using group by operations on the back end rather than doing so up front.

For further data preprocessing, we will need to convert all data (with many different data types) to numbers so we can use them for our data mining techniques. However, we have decided that this step in pre-processing will only be done at the outset of each data mining application on a case by case basis. The reason for this is that maintaining a human readable dataset will be necessary during the initial pre-processing, integration, and exploratory data analysis phases. Also, the data that will need to be converted and how may vary depending on the question, so it is better to perform that step in context. A likely example of what this will entail would be to convert the arrest attribute to binary (1s and 0s) and the description attribute to the respective encoded numeric values. When we perform this conversion, we will need to be attentive to use the correct conversion. (Manhattan distance for quantitative attributes and Hamming distance for categorical attributes). We will also need to normalize and aggregate our normalized distance matrices; although our CFNAI economic dataset is normalized and aggregated, that is not the case for our Chicago Crime dataset.

While many people have utilized the Chicago Crime dataset for data analysis and mining (please see the Literary Survey above), we did not see any past studies that integrated the crime dataset to an economic dataset. The "Chicago Crime on Christmas" is an excellent prior work that integrated the Chicago Crime dataset and a weather dataset to explore crime trends with changes in weather. It inspired us to upgrade our interesting data mining questions and add the CFNAI economic dataset as a supplement to our primary Chicago Crime dataset. Unlike previous usage of the Chicago Crime dataset, we will be exploring crime data against various economic attributes and model our data in multiple dimensions with data cubes. To gain more insights, we will also use OLAP operations to aggregate, slice and zoom into our data.

DATA SET

Our primary data set is the Chicago Crime dataset, hosted by Google's Big Query. A link to the dataset and description can be found on Kaggle, at https://www.kaggle.com/chicago/chicago-crime [9]. The data is generated and owned by the Chicago City Police. Each record represents a reported crime and includes information regarding its time, location, and various categorizations describing the crime. The dataset contains 7,213,094 entries from 2001 to 2020.

Our secondary data set is the CFNAI Historical (Real-time) data set, created and maintained by the Federal Reserve Bank of Chicago. The CFNAI metric is a monthly average of 85 series of economic data representing a wide range of economic indicators, each of which has been weighted, normalized, and adjusted for inflation. Each record represents the CFNAI metric for a given month, and includes four similar metrics each of which is derived from a subset of the 85 series whose indicators fall into one of four categories, 1) production & income, 2) employment, unemployment & hours, 3) personal consumption & housing, and 4) sales, orders & inventories. It is available for download at https://www.chicagofed.org/research/data/cfnai/historical-data.

EVALUATION METHODS

In order to evaluate the performance of our classifiers which predict economic values with our crime data, we will use metrics such as accuracy, precision and error rate. We will also perform standard evaluation techniques including R-squared, P-value, F-test and confidence intervals to observe the variance and quality of prediction. This will allow us to analyze the predictive capability of our model and make adjustments accordingly. If a model is not making predictions or classifications up to our accuracy standard, it might be in our best interest to substitute it with another model. Moreover, we can also use the holdout method and k-fold cross validation to further validate our models.

To evaluate our result for whether an arrest will be made for a commited crime, we will likely use the classification technique with a decision tree. This means that we will first use our validation test set to assess the accuracy of our classification model. Furthermore, we can access the coverage and accuracy of individual rules in our decision tree. For other correlations and associations that we want to explore between crime and economic metrics, we will perform correlation analysis and use multi level association rules to break down the underlying support and confidence.

Evidently, we will also have many visualizations to support both our EDA and data mining results. For example, plotting the root mean squared error will be helpful for us to visualize the deviation of our errors from the actual values in our crime dataset. In general, a metric that we will certainly look at is the efficiency and runtime of different methods and models.

TOOLS

The tools that will be used to explore, analyze, test and visualize our results from our Chicago Crime and CFNAI datasets will include but are not limited to: Python and its libraries including Numpy, Pandas, Matplotlib, Sklearn and Seaborn. Google BigQuery, SQL and Excel will be used to transfer our data to the python libraries. For more choices and other useful visualizations, we will use Tableau and D3.js.

MILESTONES

The most important milestones of our project from this point on will be:

1. Building an Extract-Transform-Load (ETL) pipeline
2. Exploratory Data Analysis (EDA) to determine interesting questions
3. Using data mining techniques to answer interesting questions; create report on findings
4. (Stretch goal) Development of tools for finding interesting patterns (data cube, frequent pattern analysis)

Building an Extract-Transform-Load Pipeline

Deadline: 30 October 2020

Prior to analyzing the data, it is necessary to do some initial cleaning and integration. All preprocessing will be completed in a single Jupyter Notebook, where the (remote) data is queried via SQL, converted into a Pandas DataFrame, and subsequently cleaned. Once cleaned, the DataFrame will be integrated (joined) with our econometric data set. Finally, the fully-processed DataFrame will be exported to .csv format for use in our EDA Jupyter Notebook.

Exploratory Data Analysis (EDA) to determine interesting questions

Deadline: 20 November 2020

With a cleaned and fully-integrated dataset, we can begin examining the data to find interesting trends and correlations. This process will include creating visualizations and creating summary statistics for various attributes. This process will result in fully fleshed-out questions which we will attempt to answer using data mining techniques. The final “deliverable” for this milestone will be a Jupyter Notebook including all EDA and resulting questions.

Using data mining techniques to answer interesting questions; create report on findings

Deadline: 27 November 2020

With our questions in hand, we will begin attempting to answer them using traditional data mining techniques. Some of these techniques are mentioned in the “proposed work” section above. The final deliverable for this milestone will be a Jupyter Notebook, separated by question, displaying all data mining/visualizations and answers to our questions. Given the dynamic nature of data mining, we may expand our results to multiple notebooks as warranted.

Development of tools for finding interesting patterns

Deadline: 11 December 2020

Finally, given enough time, we would like to make a generalizable tool for mining the Chicago crime dataset. This will involve building an application where users can answer their own questions about the dataset using basic data-mining techniques. The tasks that the application would support include rollup and drill down on various attributes, frequent pattern analysis, and visualization.

MILESTONES COMPLETED

In short, we have completed our first two milestones and begun work on our third. The first milestone we outlined was building our ETL (Extract-Transform-Load) pipeline, which queries the Google-hosted dataset and performs data cleaning and preprocessing, as well as integrates our supplemental datasets. The second milestone we outlined was to perform an EDA (Exploratory Data Analysis), to help ourselves and readers become more familiar with the data, for purposes of providing context, exploring initial assumptions, and for formulating more specific data mining questions. These milestones prepare us for our third and primary milestone, which is to perform the data mining techniques we have learned in class thus far in hope of gaining new information and insights.

Our first milestone was completed by creating a Jupyter notebook to run queries from, perform actions on the resulting data, and export the results for data exploration and mining. In order to access the Google-hosted dataset we each needed to create Google Cloud Platform projects with BigQuery enabled. Within these projects we each created service accounts and exported those credentials into JSON files, which can be exported into the local environment and used by the ETL Pipeline notebook to query the dataset. The notebook uses the python sql module to connect to the dataset and query all attributes and all rows of the data, yielding over seven million crime entries. It then loads the data into a pandas dataframe. From here on, we primarily use pandas for manipulating and interacting with the data. We do this by defining a series of functions to do the cleaning, preprocessing, and integration, then perform those functions on a copy of the original data. This includes all of the steps we have described in *Proposed Work*, including removing null objects, checking for duplicates, deriving separate date and time attributes as well as others, dropping irrelevant attributes, and converting certain attribute data types to data types more amenable to our chosen methods of interaction. We choose to omit null objects completely rather than fill these empty cells with mean values or other possible methods because they all occur in the location / coordinates attributes, which means most assumptions about the data would be biased. Furthermore, the occurrences are spread equally throughout the data and across different categories such that omitting them should not skew the results significantly. This reduces our dataset by only several hundred thousands out of over seven million, which to us is an acceptable tradeoff given the distribution of null values across the dataset.

Dropping uninteresting attributes such as id columns and converting the data types such as latitude / longitude into geometry objects usable by pandas is a straightforward affair, however before we can join the CFNAI economy data we needed to derive a “month” attribute from the date column to join it on. We also needed to read in the CFNAI data and convert its data column to a similar format. While creating the derived month attribute and joining the CFNAI data onto it does create many duplicates of the CFNAI data, at this point in our process that is entirely acceptable because our goal is not to create the most efficient schema for later data mining purposes, but instead to form a “data lake” that aggregates all of our data and supplemental data into a single dataframe that can later be copied and then decomposed or sliced into whatever subsets we require for a particular data mining task. Pandas dataframes are a great tool for this purpose. After joining the econ data, we then proceed to join the crime type categories data along the fbi_code column. We originally intended to join this along the ucr column, which would have provided another facet to analyze the data on, however it turns out that the Chicago Police have mislabeled the codes they are using and in fact have not finished converting their data to the new NIBRS Uniform Crime Reporting standards, and the fbi_code attribute actually still conforms to the old SRS crime reporting standards while their ucr column has been updated. Nonetheless by joining this data on the fbi_code column we were still able to derive new attributes for each object including the fbi_code category, index (more serious) crime, violent crime, and whether the crime is against a person, property, or society. Once these steps are applied to a copy of the stored dataframe, the ETL Pipeline exports the cleaned data into a csv file. The entire dataset results in a 2.75GB file, which is substantially less than we originally anticipated. This is ideal in that it allows us to transfer the cleaned data to each user who can then then load the data into other notebooks for exploration and data mining.

Our second milestone, performing the exploratory data analysis, was also done within the context of a Jupyter notebook. We read the cleaned data in from our csv file, created by the ETL pipeline, into a pandas dataframe. From there, we were able to perform a thorough overview analysis of the primary data and its relation to the integrated data using basic functions and by plotting those results using the seaborn vis library. We leave a detailed description of our findings there until later in our report, under the *Results So Far* section.

MILESTONES TODO

With the groundwork laid, we are on schedule to complete the next milestone in our project. Namely, finally applying the data mining techniques we have learned in class so far. We will be initially oriented by the questions we have posed up front in our problem statement as well as by the questions raised during our exploratory data analysis (described in the EDA notebook). Initial looks at the correlation between economic and crime data were negative, however this has the potential to be even more interesting than it would have been otherwise if by using data mining techniques we are able to reveal any correlations / associations between subsets of the crime data and / or econ data. Individual questions and methods will be explored by group members either individually or through pair programming sessions using individual Jupyter notebooks. Any additional preprocessing steps required specific to the question at hand will be applied at this point, for example data aggregation, data reduction, or converting certain categorical values into numeric form. As new questions arise or new methods are warranted for a given question, new notebooks will be spawned in order to promote more collaborative input from all members unhindered by the need to coordinate commits and pull to the git repository. With our finding in hand, we will consolidate information and insights gained into one or more Jupyter notebooks designed for presentation purposes. We will convert the finding back into human interpretable form by providing labels to our results and utilize data visualizations tools to explain those results. This will include either plotting results directly using seaborn and matplotlib, or by exporting the results to a csv file and using Tableau to create dashboards that can be exported to PDF format and then loaded into the notebook for final presentation.

As mentioned, Milestone 4 will be optional and is contingent both on the perceived success of Milestone 3 and the time it takes to complete. Given the nuance of factors contributing to crime not just in Chicago but in any city described at the outset of our report, we believe that even more beneficial than sharing our findings would be to share a repeatable format for continuing to explore and find interesting patterns beyond what we have done here. We would like then to formalize our process generally, but specific to our dataset although potentially including others as well, with the goal of providing these tools in a way that would allow others to skip some of the setup and proceed directly to applying data mining techniques they may be considering that we have not thought of or had time to explore. The generalized version of our project tooling along with the requirements and instructions would likely be published in a subfolder of our github repository where it would be available for anyone to access.

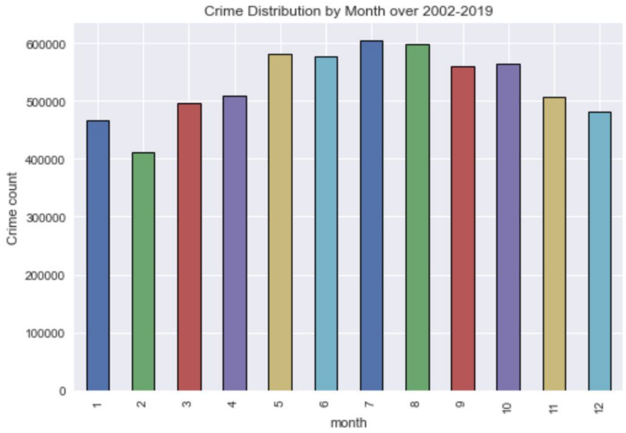
RESULTS SO FAR

So far, we have completed the EDA (Exploratory Data Analysis) and started on our data mining process. EDA helped us better understand our dataset to observe many general Chicago crime patterns. It is also an imperative process that allows us to fine tune our data mining questions even more and start making comparisons between models for our data mining techniques. Since data visualization is a core component of EDA, we also analyzed our dataset carefully with many different charts and graphs.

Exploring all of our cleaned data and dropping objects from year 2020 (it might skew our results since year 2020 is not over yet), we first utilized a breadth first approach to answer many questions involving the most important attributes that we will be using in our data mining process. First of all, “what are the most frequent and least frequent types of crime?” We found that theft and battery are two crime types that are overwhelming prevalent while human trafficking and ritualism have occured the least over the years. Diving into our dataset for the most frequent and least frequent locations of crime, we found that streets and residences contain the highest amount of crime at 1,633,671 and 1,056,497 counts respectively. On the other hand, junk yards and banquet halls are just some of the places with least occurrences of crimes.

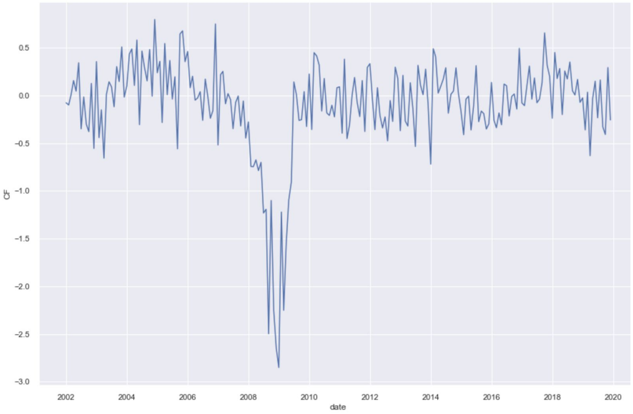
Next, we found the number of arrests that were made from 2002-2019 to be 1,738,417. Knowing that there are a total of 6,365,627 criminals in our cleaned dataset, this means that only ~27% of criminals have been arrested so far. This was a crucial initial finding since we have a data mining question on the prediction of whether an arrest will be made for a commited crime. To aid our data mining for different crime properties and associations, we filtered for the number of domestic crimes and its ratio to non-domestic crimes. Interestingly, non-domestic crimes are about 6.5 times more likely to occur than domestic crimes.

To analyze the count and changes of crime over time, we plotted “crime vs. year”, “crime vs. month” and “crime vs. hour”. Our resulting bar plot indicated that since 2003, crime has been steadily dropping before bottoming out in 2015. From 2015 to 2019, crime count has been fairly close every year. Zooming into monthly crime counts, it is clear that crime peaks in months of July and August. It is also interesting to note that the total number of crimes is much lower in February relative to any other month.



In terms of crime trends over hourly increments, we found that the highest number of crimes occurred mid-day at about lunch time. This is a surprising result since we didn't expect any crime peaks to occur during the day. However, as we predicted, later hours do generally have more occurrences of crimes when compared to the early times in the morning when people wake up. To analyze patterns of different crime types, we derived several more attributes including "index crime", "violent crime" and "property crime." Interestingly, our bar plots show that the majority of crimes are non-violent.

To understand our economic data, we explored the correlations between our economic attributes. If we are able to find a strong correlation between two economic attributes, we will be able to use just one to find correlations with crime and eliminate the other extraneous attribute. Plotting all economic attributes against each other, we noticed that many of them have clear positive correlations. However, we also saw outliers that come primarily from the year of 2008. Using boxplot and time series, we were able to verify that the sharp drop in our economic measures corresponds to the Great Recession.



CF attribute dives from 2007 to 2009

To explore our data mining questions about the relationship between crime and the economy, we plotted total crime against our economic metrics and didn't find any noticeable correlations. Interested in how total crime and the CF measure change over time, we also plotted them on the same graph. The result indicated that while total crime doesn't appear to dip with the economy during the 2008 recession, there is a significant dip for both total crime and CF at the start of COVID in 2020.

REFERENCES

[1] Mary Pattilo. 2018. Crime in Chicago: What Does the Research Tell Us? Retrieved October 2020 from https://www.ipr.northwestern.edu/news/2018/crime-in-chicago-research.html.

[2] Matt Ford. 2017. What's Causing Chicago's Crime Spike? *The Atlantic*. Retrieved from https://www.theatlantic.com/politics/archive/2017/01/chicago-homicide-spike-2016/514331/.

[3] Kaggle. 2019. Chicago Crime on Christmas. Retrieved October 2020 from https://www.kaggle.com/mkrkkinen/chicago-crime-on-christmas.

[4] Cjango. Chicago Crime Data Analysis (Python Project). Retrieved October 2020 from https://cjango.wordpress.com/portfolio/chicago-crime-data-analysis-python-project/.

[5] Kaggle. 2019. Chicago Codes - Crime Map and BigQuery in R. Retrieved October 2020 from https://www.kaggle.com/alkadri/chicago-codes-crime-map-and-bigquery-in-r.

[6] Sadaf Tafazoli. 2018. My Notes on Chicago Crime Data Analysis. Retrieved October 2020 from https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis-ed66915dbb20.

[7] Julia Burdick-Will. 2013. School Violent Crime and Academic Achievement in Chicago. *Sociology of Education.* PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831577/.

[8] Chicago Police. Crime Type Categories: Definition and Description. Retrieved from http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.

[9] Kaggle, Chicago City Police. 2018. Chicago Crime (BigQuery Dataset). Retrieved October 2020 from https://www.kaggle.com/chicago/chicago-crime/metadata.

[10] Federal Reserve Bank of Chicago. 2020. CFNAI Historical (Real-Time) Data. Retrieved October 2020 from https://www.chicagofed.org/research/data/cfnai/historical-data.