

Machine Learning for Predicting Ward Success in Chicago

Olivia Fenwick, Miguel Gutierrez,
Meghan Rokas, and Aliya Zhdanov

Agenda

Problem Statement and Research Objectives

Exploratory Data Analysis

Data Preparation

Models and Insights

Lessons Learned and Recommendations

Our Team



Olivia Fenwick

FT Student, 3rd Quarter
B.S. in Mathematics and
Economics from University
of Michigan



Miguel Gutierrez

FT Student, 3rd Quarter
B.S. in Astrophysics and
Mathematics from Florida Tech




Meghan Rokas

FT Student, 3rd Quarter
B.S. in Biology from
Loyola University Chicago



Aliya Zhdanov

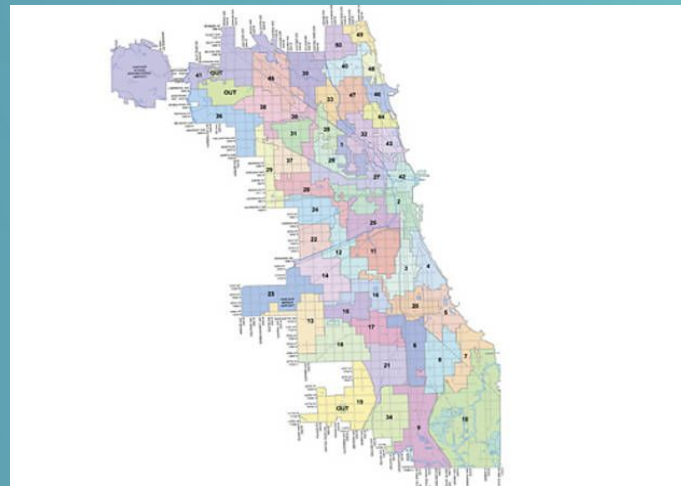
FT Student, 3rd Quarter
BS in Business & Statistics
from Carnegie Mellon
University



Problem Statement and Research Objectives

Executive Summary

- Chicago is segmented into 50 wards
- Each ward has its own alderman who acts as a liaison between their community and the greater Chicago area
- The fifty aldermen together make up the city council which is responsible for making decisions regarding things like healthcare, welfare, and taxation
- Ward health can be categorized by many things: happiness of its residents, crime rate, number of new residents moving into the neighborhood, number of businesses, etc.
- We will focus on crime numbers and business licenses for the scope of this project



Problem Statement

- Contributions/gifts made to aldermen total tens to even hundreds of thousands of dollars per year
- Where is this money going?
- We will analyze a variety of datasets that pertain to Chicago's 50 wards
- Specifically interested in whether there is a correlation between ward health and the contributions made to the aldermen
 - Are the wards with less crime, more businesses, etc. also the ones getting a lot of donations?
 - Is ward health predictive of the amount of gifts received?



Ald. Brendan Reilly, Ward 42, Downtown Chicago



Ald. Leslie Hairston, Ward 5, Hyde Park, Chicago

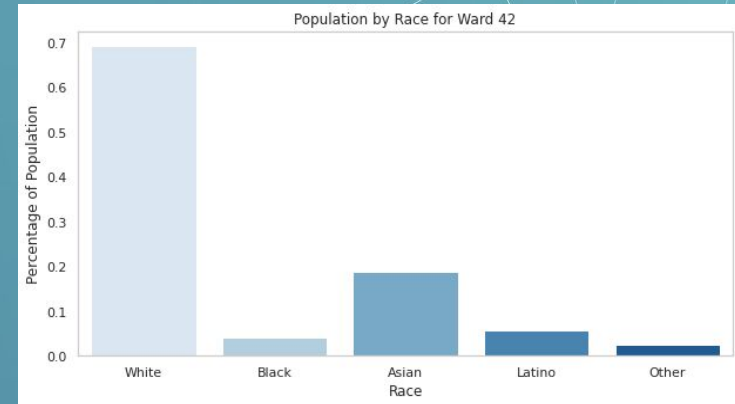
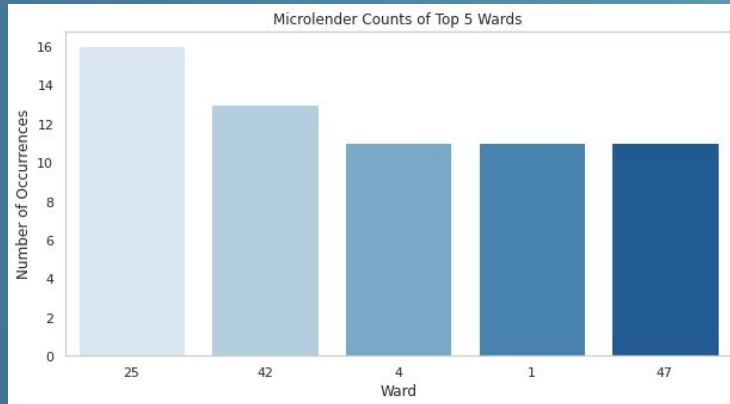
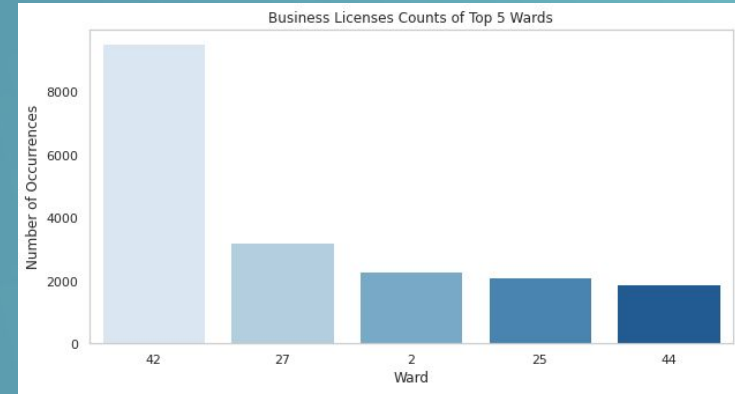
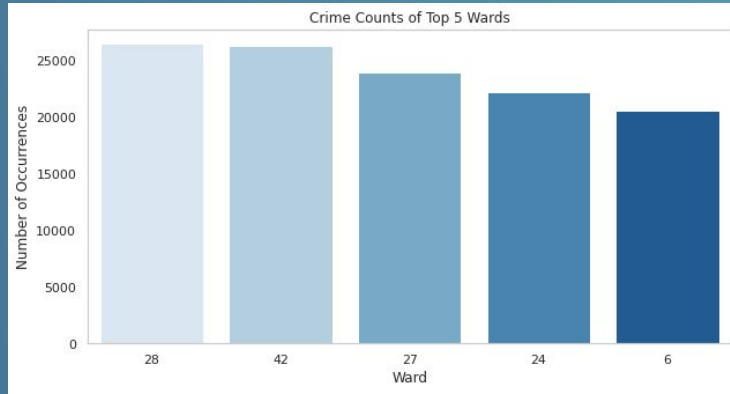


Exploratory Data Analysis

We used many data sources from the City of Chicago data bank

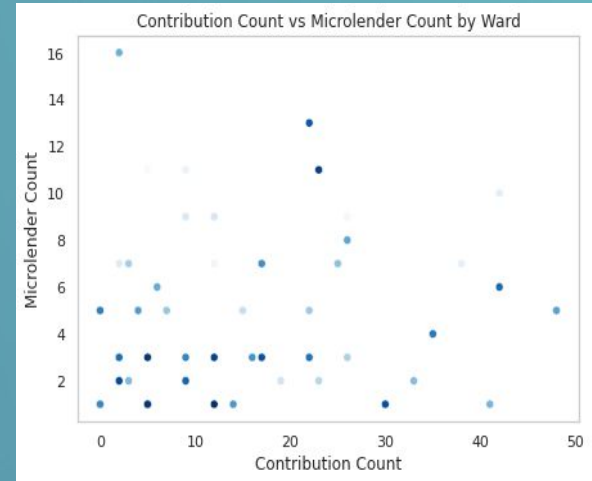
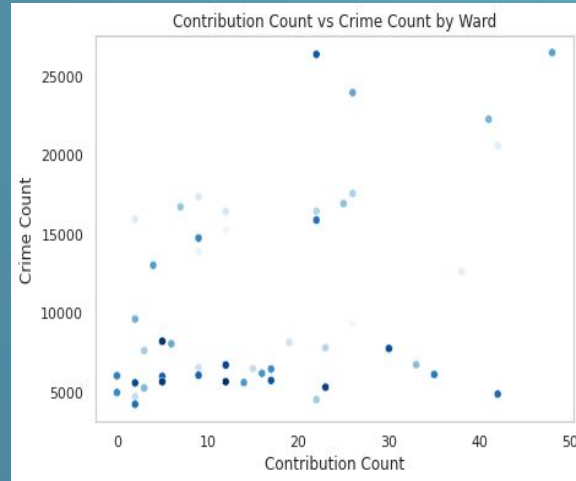
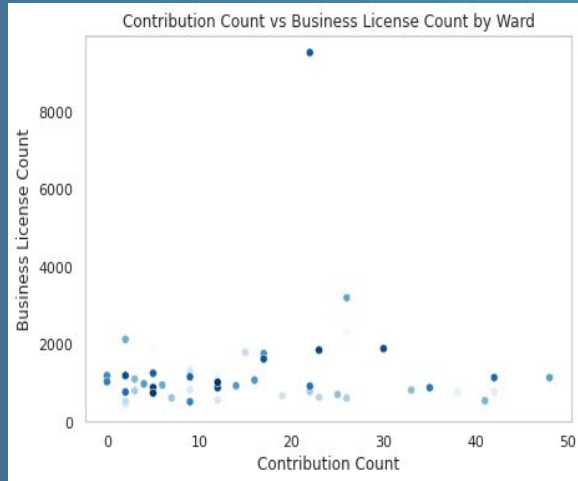
Data	Size of Dataset	Number of Observations	Number of Attributes
Lobbyist Data - Gifts	106 KB	848	13
Lobbyist Data - Employers	709 KB	5677	13
Lobbyist Data - Compensation	3 MB	28,958	11
Lobbyist Data - Lobbyists by Client	2 MB	5857	7
Ward Offices	12 KB	50	16
Business Licenses - Current	24 MB	70,671	34
Crimes 2001 - Present	2 GB	7,333,802	22
Microlending	32 KB	252	14
Estimated Ward Populations	23 KB	848	13

Ward 42 (downtown) is in the top 5 for each of our datasets



Crimes data appears to have a potential relationship with contributions data

- On initial inspections it appears that total contributions from lobbyists has a positive correlation with total crimes for some wards
- However, it seems like there's no relation with business licenses or microlenders per ward





Data Preparation

Despite only having only 50 rows, our data generated 90 columns through feature engineering

- Used data from 2019 onwards since that was the last alderman election and ensured complete and consistent data for contributions
- Original data had a row for each instance of a business license/criminal activity/microlender/contribution and we generated our columns from that data
 - Rolled up business license/criminal activity/microlender/contribution to get total count
 - Calculated summary statistics for contribution amount (mean, median, total sum)
 - A total count column for each type of crime/business license



Many of the columns in the data are highly correlated

- Performed PCA to reduce multicollinearity
 - Many columns such as weapons violation and criminal damage are highly correlated
 - Scale data to normalise before PCA
 - Reduced from 90 columns to 21



The overlap in contributions data was measured and used to create connected graphs.

- Done using a modified version of an example on GitHub
- 3 Different Networks:
 - Wards connected by lobbyists
 - Recipients connected by lobbyists
 - Lobbyists connected by clients

```
In [56]: 1 ovlp_dict
```

```
Out[56]: {'ALDERMAN MARIA HADDEN': {},  
          '24TH WARD ORGANIZATION - ALD. MICHAEL SCOTT': {},  
          'ALD. GILBERT VILLEGAS': {},  
          'JASON ERVIN': {'FRIENDS OF WALTER BURNETT': 3,  
                          'FRIENDS OF LESLIE HAIRSTON': 3,  
                          'FRIENDS OF ED BURKE': 3,  
                          'CARRIE AUSTIN': 3},  
          'VILLEGAS FOR COMMITTEEPERSON': {'HOPKINS FOR CHICAGO': 3},  
          'FRIENDS OF MICHAEL RODRIGUEZ': {'FRIENDS OF GILBERT VILLEGAS': 3,  
                                             'HOPKINS FOR CHICAGO': 3,  
                                             'CITIZENS FOR ERVIN': 3,  
                                             'CITIZENS FOR WAGUESPACK': 4,  
                                             'CITIZENS FOR JOE MOORE': 3},  
          'FRIENDS OF GILBERT VILLEGAS': {'FRIENDS OF RODERICK SAWYER': 5,  
                                             'FRIENDS OF MICHELLE A. HARRIS': 3,  
                                             'FRIENDS FOR DEBRA SILVERSTEIN': 5,  
                                             'FRIENDS OF SUSAN SADLOWSKI GARZA': 4,  
                                             'HOPKINS FOR CHICAGO': 11,  
                                             'CITIZENS FOR ERVIN': 6,  
                                             'FRIENDS OF WALTER BURNETT': 3}
```



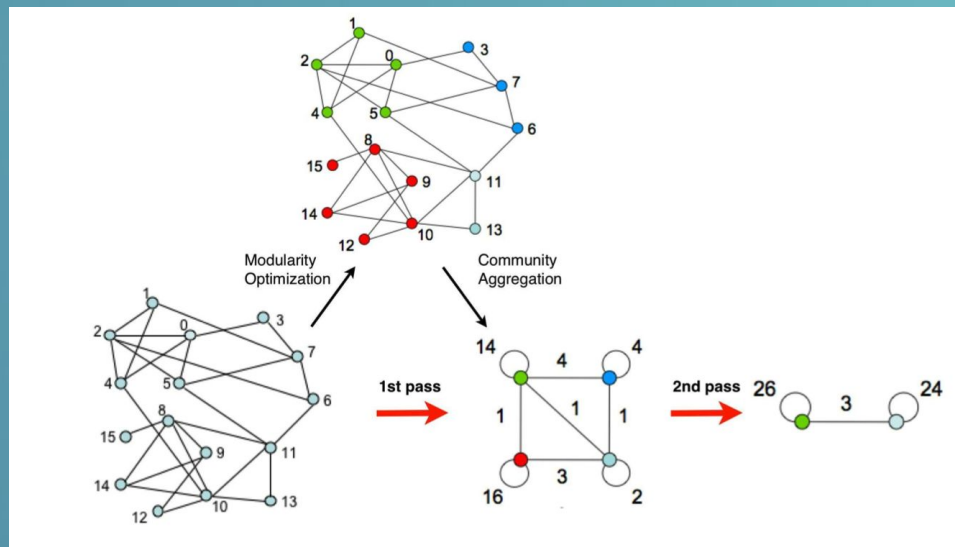

Models and Insights

Using a network representation of contributions that tracks overlaps, communities can be discovered.

- The Visualization platform *Gephi* implements an algorithm developed by Blondel et. al. that maximizes an objective function that targets modularity to detect communities within a network
- Modularity is defined as:

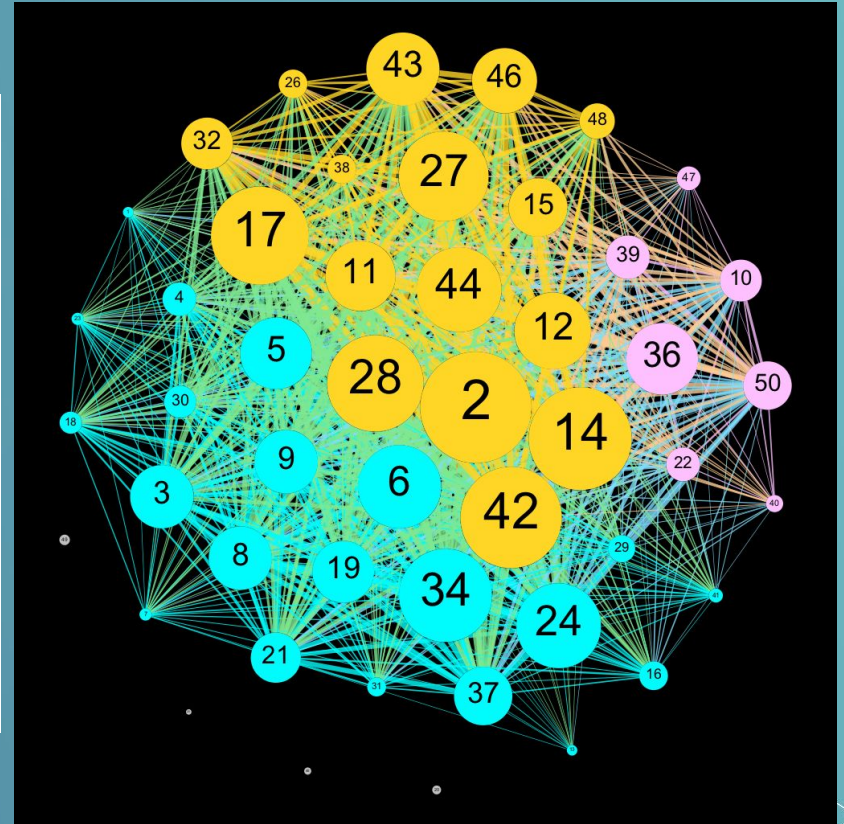
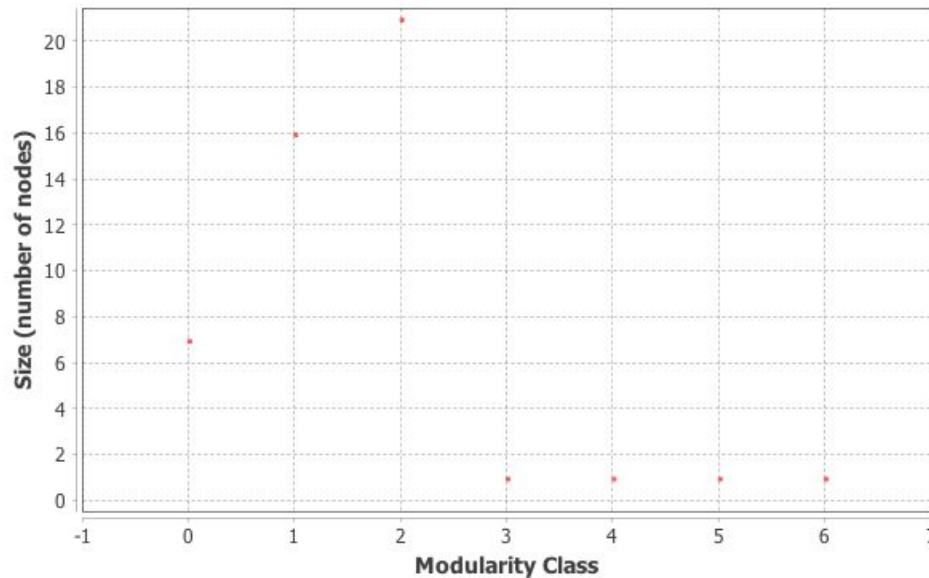
$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

Where A_{ij} is the weight of the edge between nodes i and j , k_i is the sum of the weights of all edges connected to i , c_i is the assigned community and m is $\frac{1}{2}$ the sum of all of the edges A_{ij} .



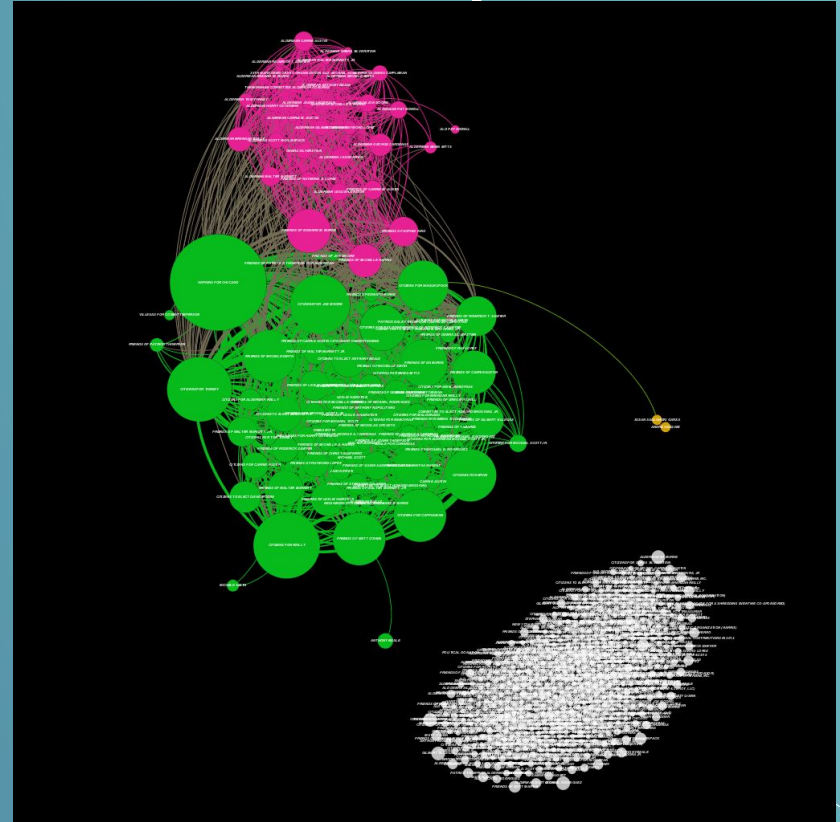
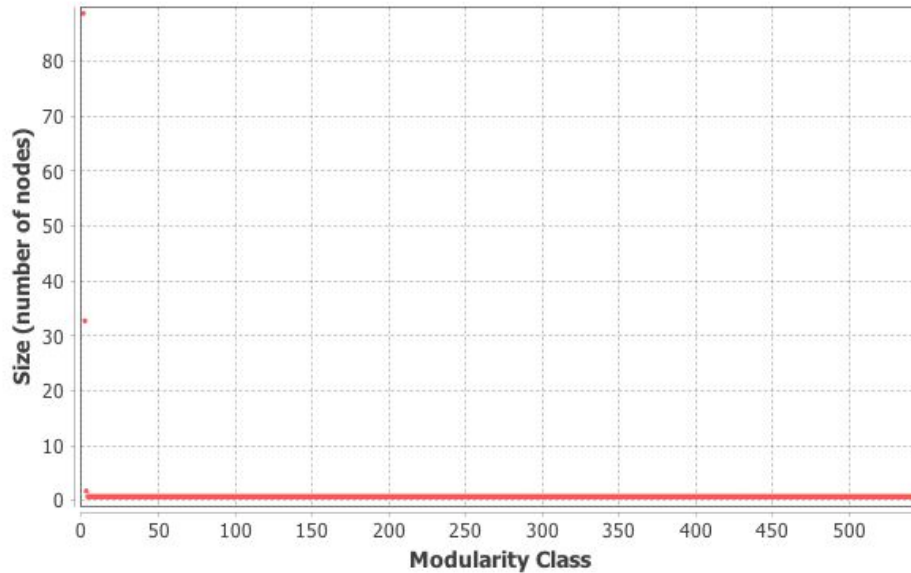
A modularity analysis of the network of wards and lobbyists found 3 distinct communities of wards.

Size Distribution



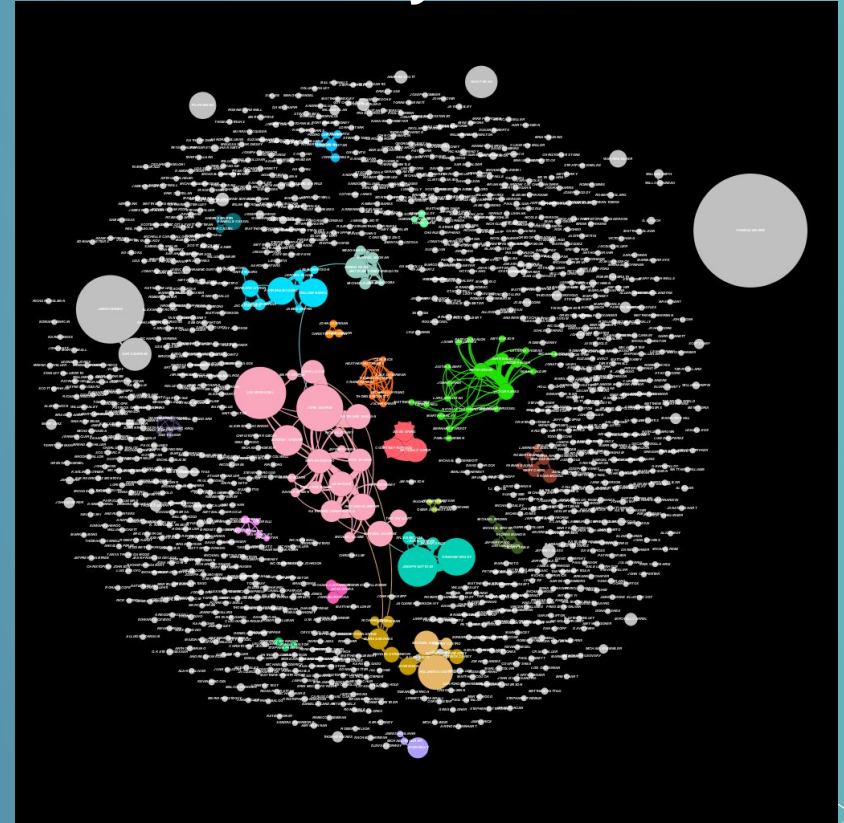
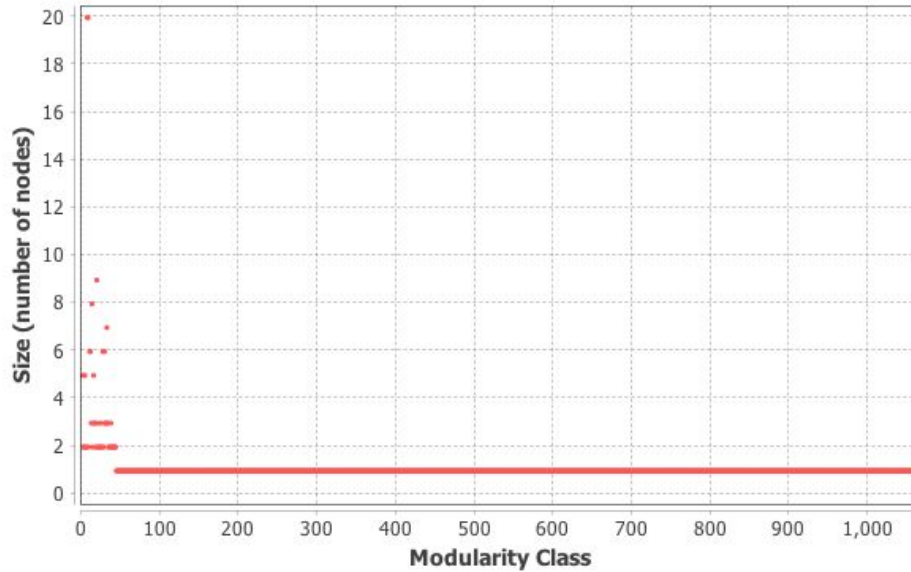
A modularity analysis of the network of recipients and lobbyists found 3 distinct communities of recipients.

Size Distribution



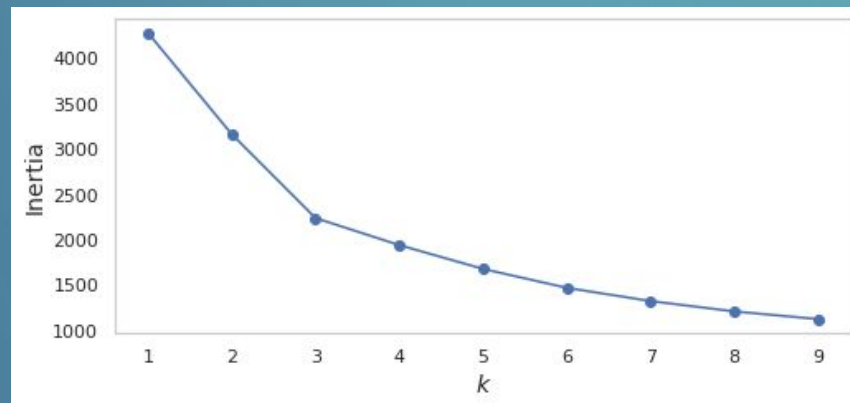
A modularity analysis of the network of lobbyists and companies found 21 communities of 3+ lobbyists.

Size Distribution

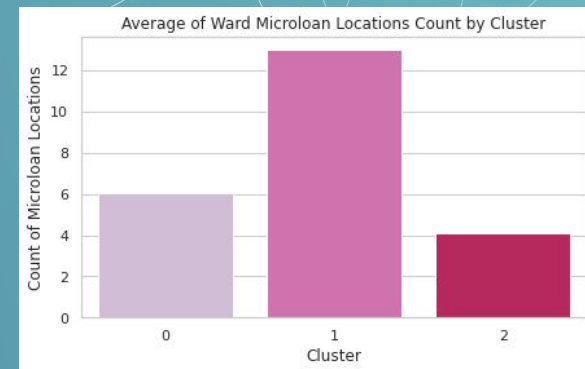
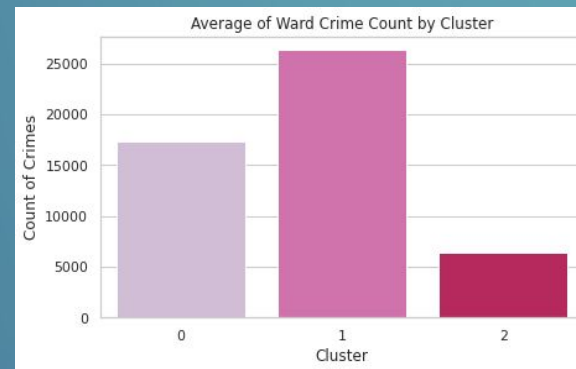
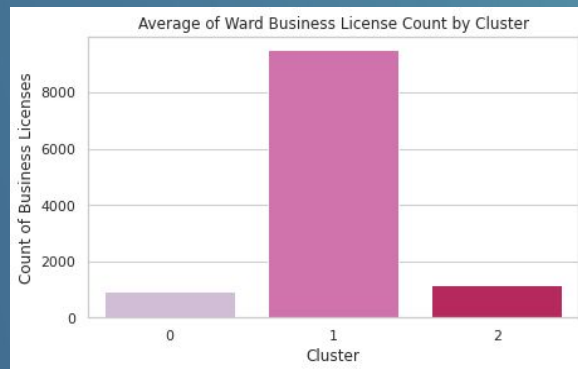
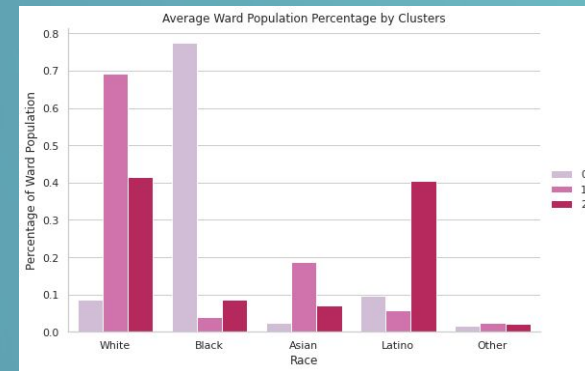
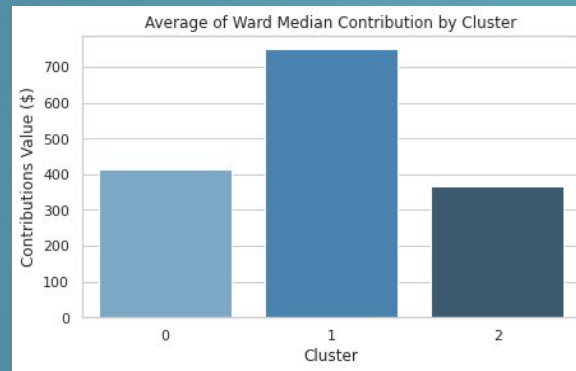
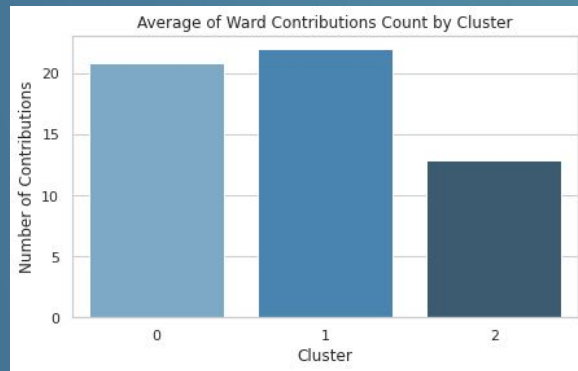


K-means revealed three distinct clusters of wards

- Trained the model on business license/crime/microlender/population columns and held out contributions columns for analysis
- Used elbow method to identify correct number of clusters in the data
 - Three clusters were identified
- After training the model one of the clusters only contained Ward 42 (downtown Chicago)



Ultimately geography played the largest difference in the clusters



We also built several models to predict contributions across wards

Independent Variables

Crime Rates
Demographics
Business Licenses

Independent Variables

Number of Contributions
(2019-2021)

Pipeline

Test/Train Split

Standard Scaling

PCA (95% retained)

Model Build

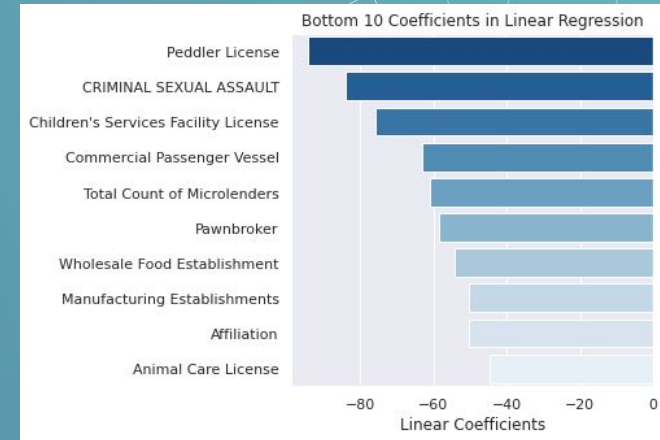
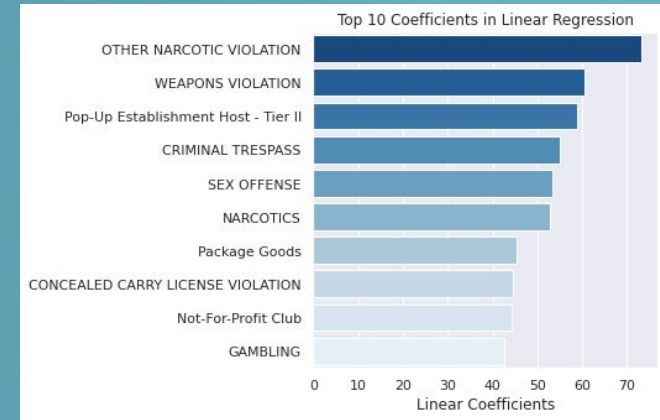
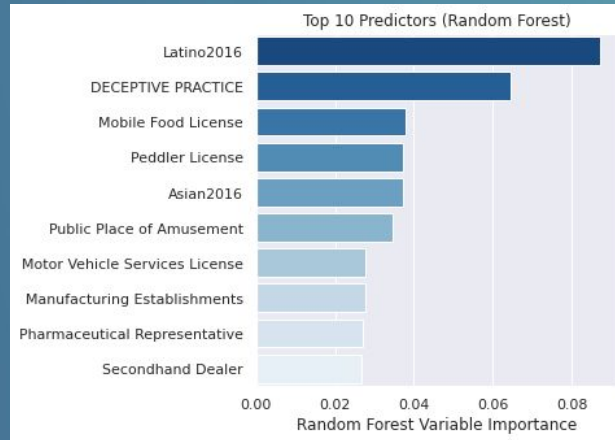
Predictive performance was poor for test data

Model	Train RMSE	Test RMSE	Train R2	Test R2
Linear	1.36×10^{-12}	39.312	1.0	-5.78
Lasso with PCA	122.907	303.594	0.703	-5.5435
Ridge	0.016305	39.1178	0.9999	-5.7135
Random Forest	4.2165	14.7421	0.8589	-5.7135
Ensemble with PCA	62.41496	239.095	0.923	-3.058
Gradient Boost	0.0612	16.6649	1.0000	-0.2184
ADA Boost	2.1285	16.0591	0.9640	-0.1315
Sequential Deep Neural Network	9.4263	31.1949	-13.2425	-2.5002

Variable Importance

Running a simple linear regression model on scaled, non-PCA data allows us to see potential major factors in ward contributions

- Wards with high narcotics & weapons crimes tend to have more contributions
- Wards with more peddler and animal care licenses tend to have fewer contributions



Our findings could be used by Chicago residents to keep aldermen accountable and track lobbyist activity

- If used by the city - the lobbyist groupings (KMeans & modularity) could be published out to the Chicago Data Portal and visualized using Plot.ly
- If used by residents - the data could be visualized using Tableau
- Daily data refresh possible given the input data resource refresh timing



**CHICAGO
DATA PORTAL**





Lessons Learned and Recommendations

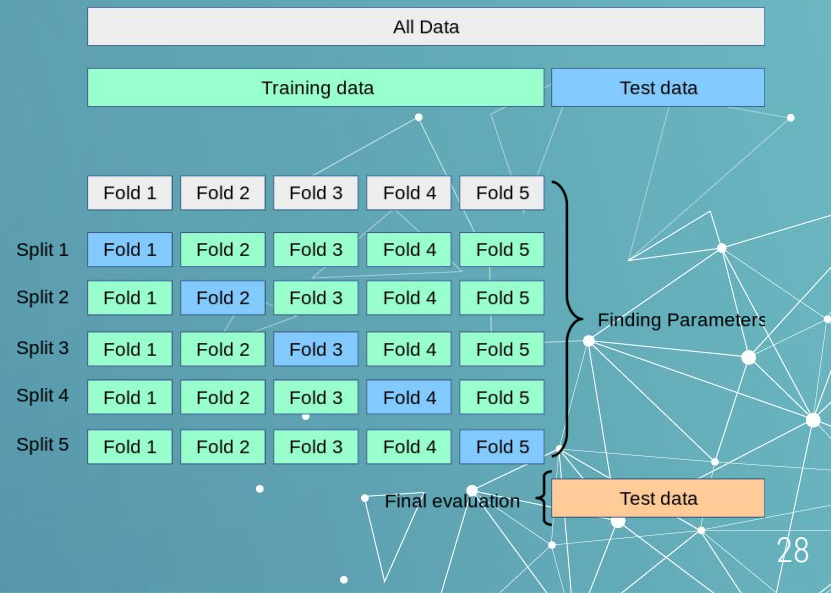
Lessons Learned

- Having a lot of data does not guarantee a successful outcome
 - Geographic data is difficult to join without the right ID column
- Data on the Ward-level is not feasible because 50 data rows is not enough, no matter how many predictors we use
- Choose your target variable wisely
 - Direction of causation matters - does corruption lead to ward failures, or do ward failures lead to corruption?
- Track target leakage
 - Suspiciously high performance could indicate that your underlying variables already contain the answer you are looking for



Recommendations

- Convert this to a different geographic grain
 - US Census blocks could help get more granular and have more records to work with
 - Wards overlap across census blocks, zip codes, and community areas
- Further text mining to identify additional benefactors
- Perform analysis across time to generate more data
- Other models/methods to try
 - Bootstrapping for all models
 - Cross validation
 - SMOTE-R





Appendix

Literature Review

Article/Website	Data	Reference	Quick Takeaways
Chicago Data Portal	X		data on: lobbyists, ward offices, business licenses, crimes, and microlending
Chicago Data Guy	X		data on ward demographics
Visualizing Twitch Communities - GitHub		X	visual mapping reference that we based our lobbyist mapping visualization on
Geron Fundamentals of Machine Learning		X	referenced geron notebooks throughout the project for various machine learning algorithms
Neural Networks for Regression - Towards Data Science		X	looked outside of geron for a specific article that discussed neural networks for regression

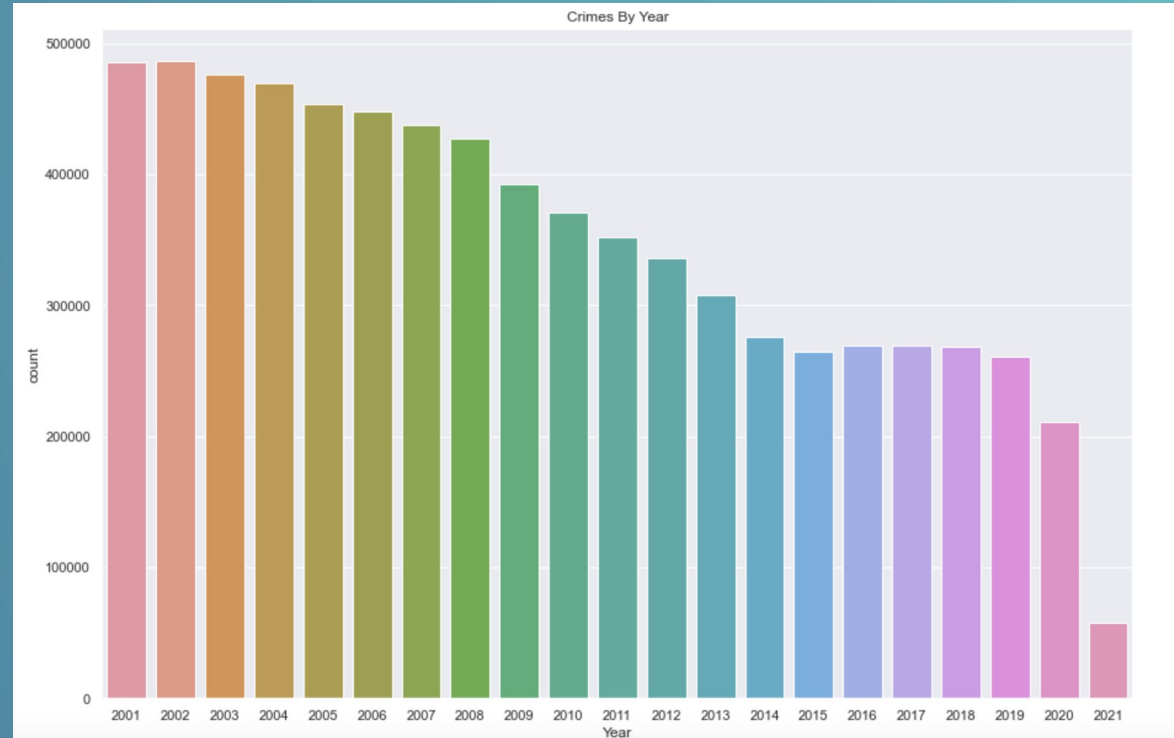
References

Datasets

- Lobbyist Data - Gifts
 - <https://data.cityofchicago.org/Ethics/Lobbyist-Data-Gifts/uzvr-cwfr>
- Lobbyist Data - Employers
 - <https://data.cityofchicago.org/Ethics/Lobbyist-Data-Employers/94gr-m4gv>
- Lobbyist Data - Compensation
 - <https://data.cityofchicago.org/Ethics/Lobbyist-Data-Compensation/dw2f-w78u>
- Lobbyist Data - Lobbyists by Client
 - <https://data.cityofchicago.org/Ethics/Lobbyist-Data-Lobbyists-by-Client/pvu3-9dfs>
- Ward Offices
 - <https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Ward-Offices/htai-wnw4>
- Business Licenses - Current
 - <https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses-Current-Active/uupf-x98g>
- Crimes 2001 - Present
 - <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2>
- Microlending
 - <https://data.cityofchicago.org/Community-Economic-Development/Chicago-Microlending-Institute-CMI-Microloans/dpkg-upyz/>
- Estimated Ward Populations
 - <http://robparal.blogspot.com/2019/01/updated-chicago-ward-population.html>

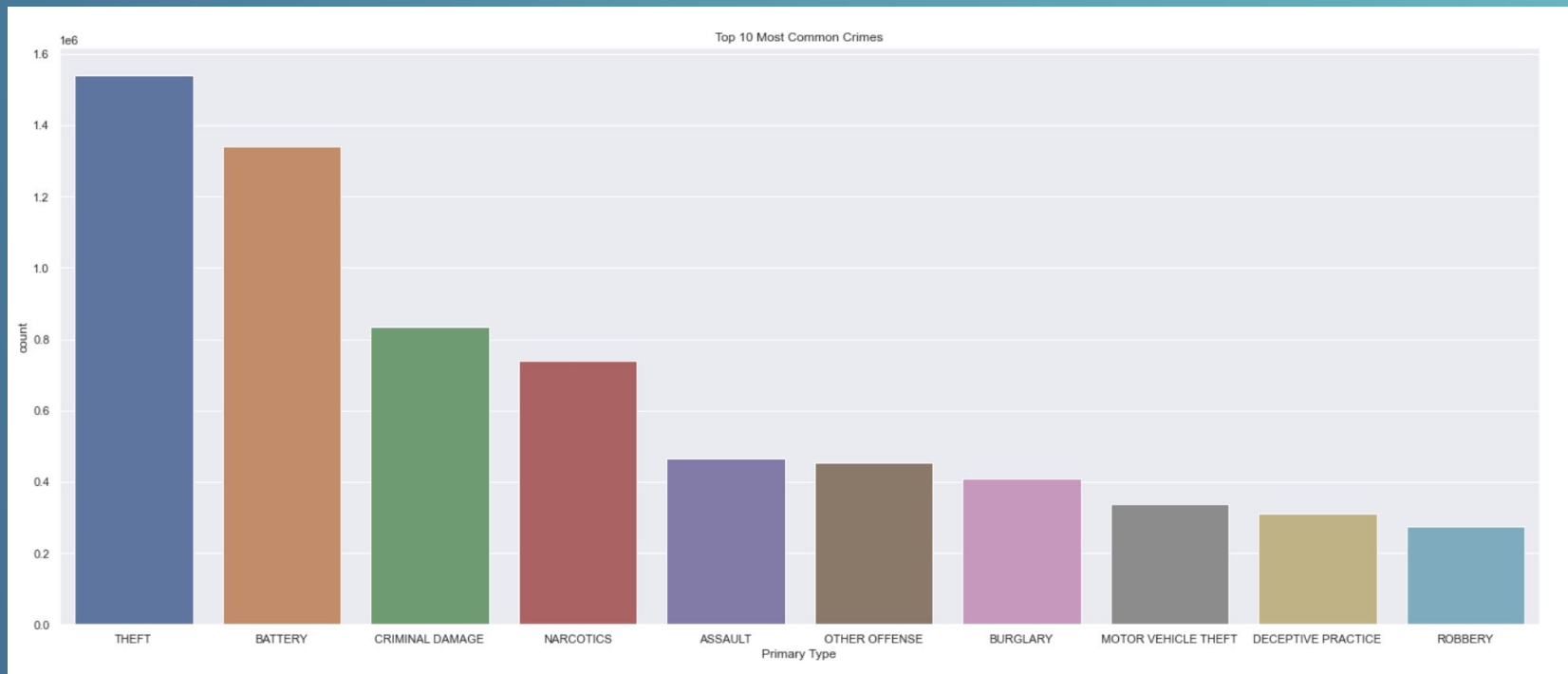
Crimes Data - Crimes Committed by Year

- This visual represents the total number of crimes per year in Chicago
- We will discount the year 2021 because we are only in May, and cannot accurately account for this year's crimes
- As you can see, there is a downward trend in crimes
- But, we have to consider the possibility of underreporting crimes in recent years
- For the sake of this project, we will assume that this data is complete and accurate



Crimes Data - Most Common Crimes

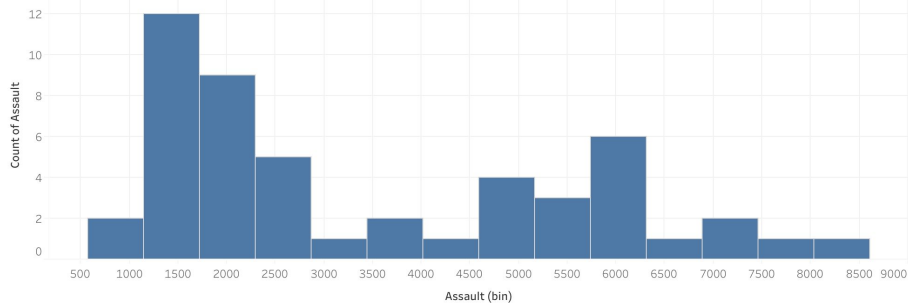
- Histogram of the most common crimes since 2001
- Theft and battery represent the most common crimes, with about 1.5 million and 1.3 million cases respectively



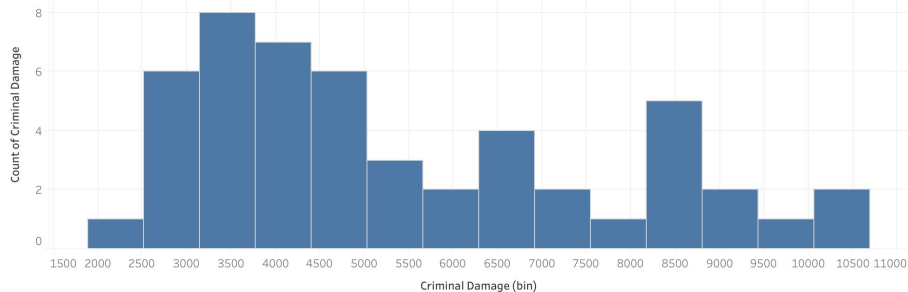
Crimes Data - Distribution by Ward

- Histogram of total Assaults and Criminal Damage across all 50 wards.
- Map of total counts of crime per ward. Seeing as Ward 42 (The Loop) is the wards with the highest count, perhaps we should bring in census data and normalize by population.

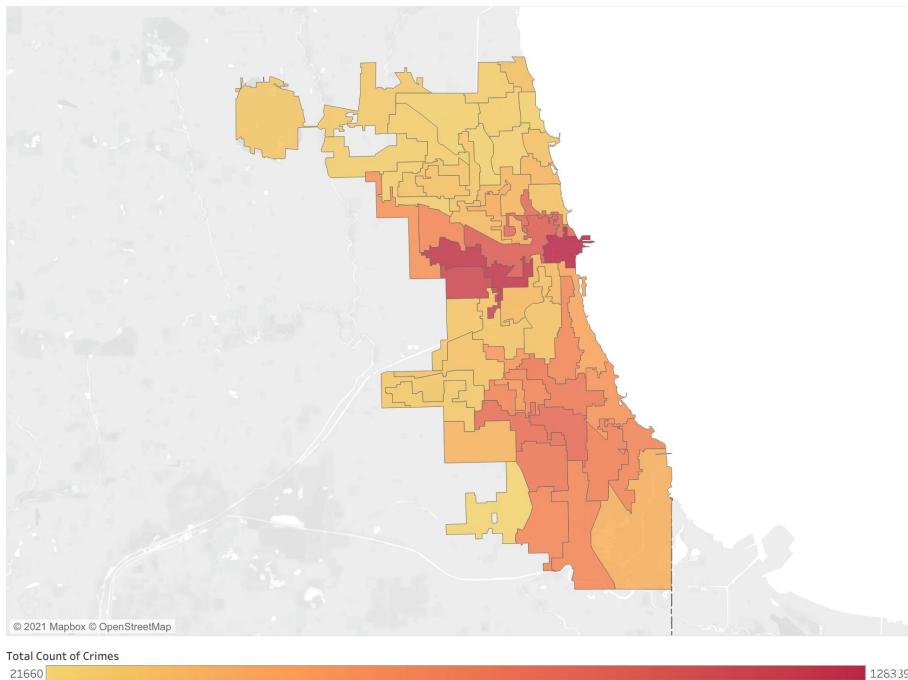
Counts of Assault



Counts of Criminal Damage

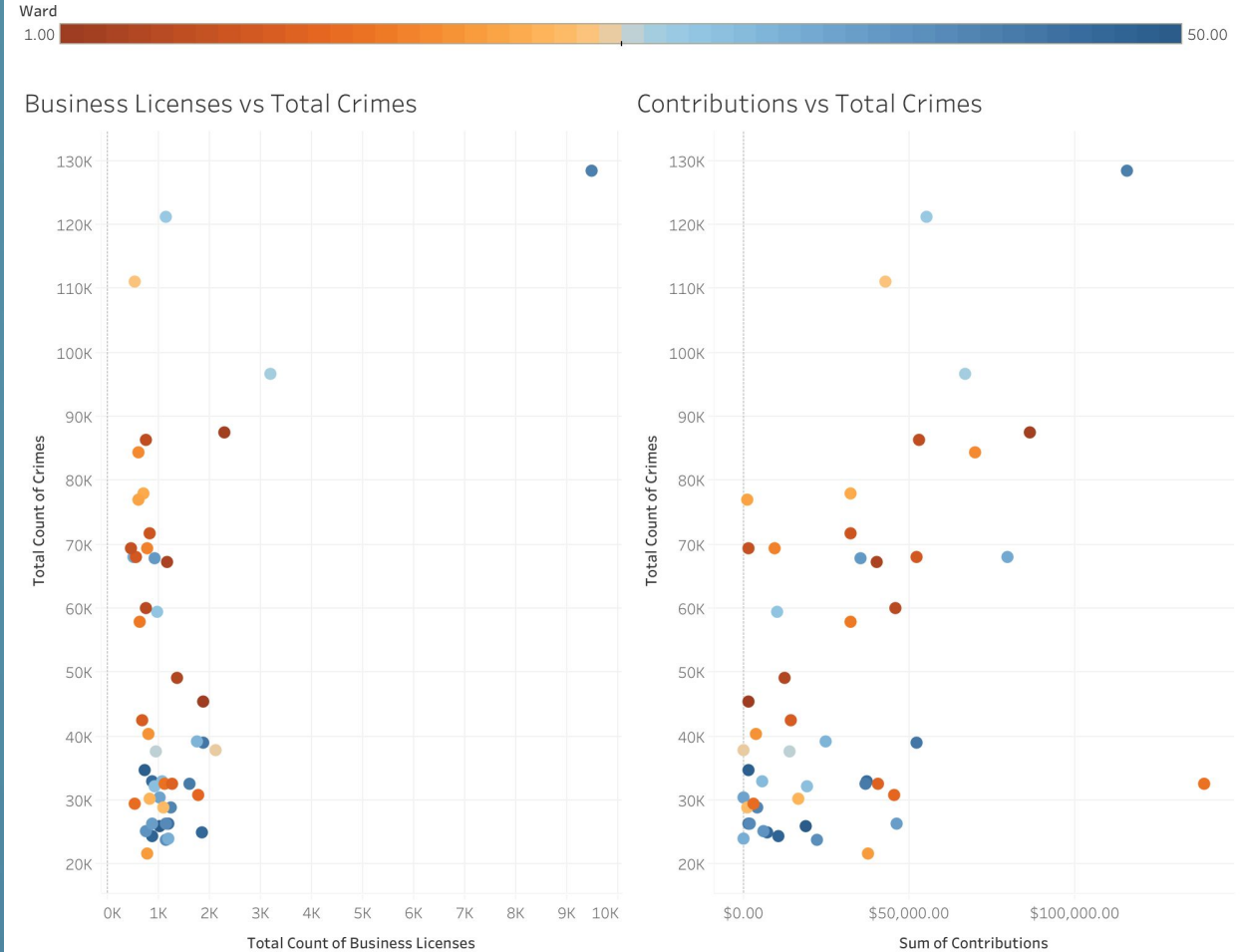


Total Counts of Crime per Ward



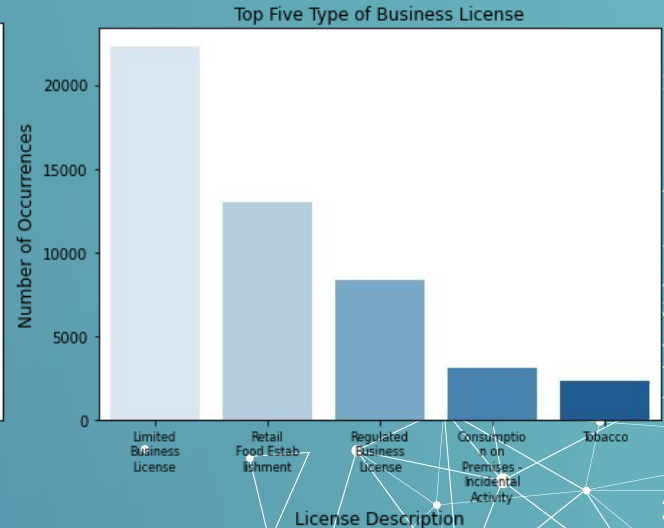
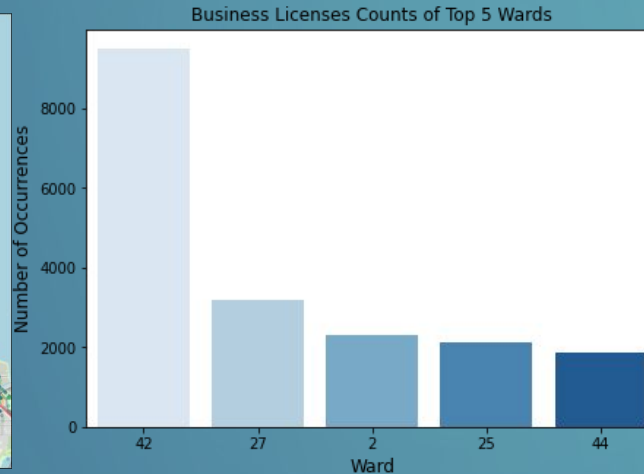
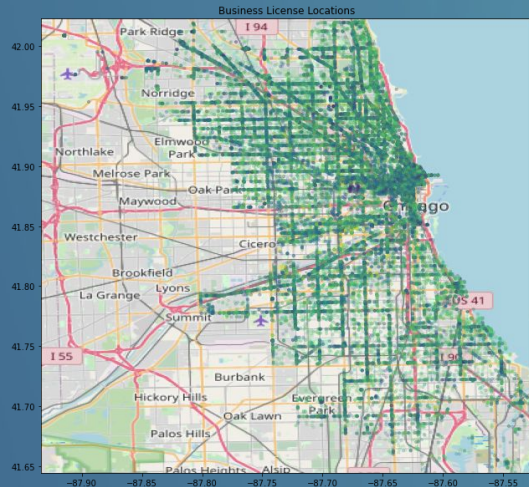
Crimes Data - Comparison to Licensing Data & Contributions

- At a glance, it appears that total contributions from lobbyists has a positive correlation with Total Crimes in a ward, while Business Licenses does not. This should be further investigated and normalized by population

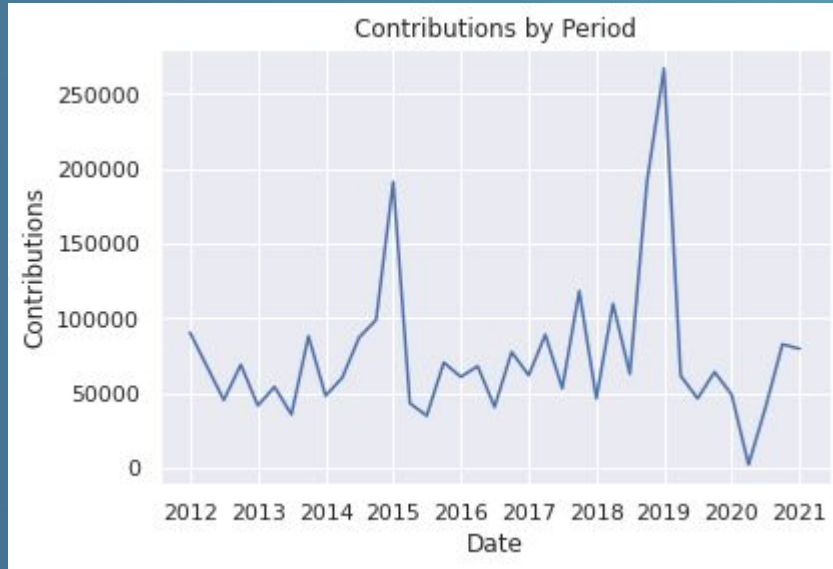


EDA of Business License Data

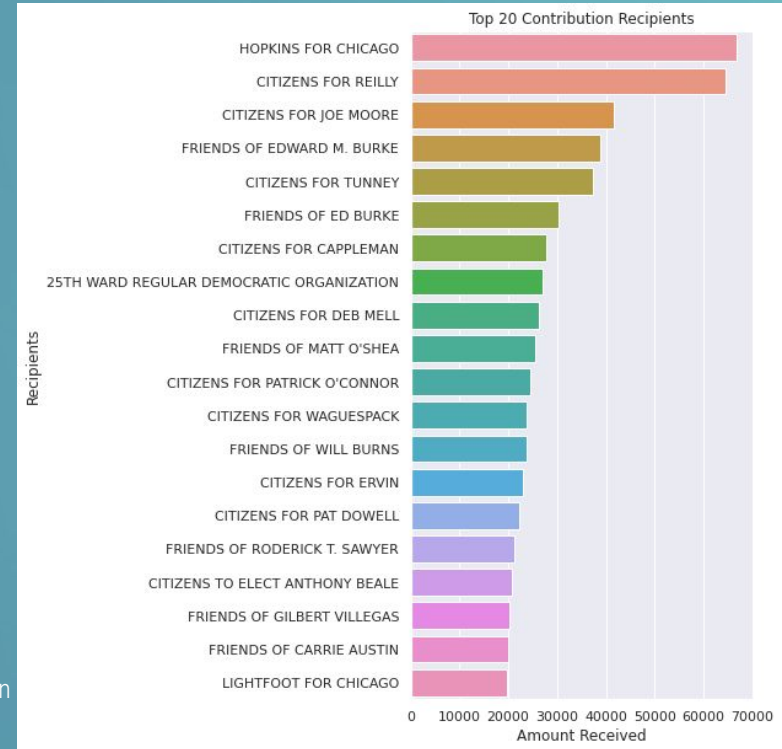
- 15.08% of the businesses with licenses granted are located in Ward 42, the area surrounding Navy Pier in downtown Chicago
 - Nearly 72.67% of licenses granted are located north of latitude line 41.85
- There are 48 different possible license types issued
 - 77.86% of business licenses issued were limited business, retail food establishment, regulated, consumption on premises - incidental activity, or tobacco licenses



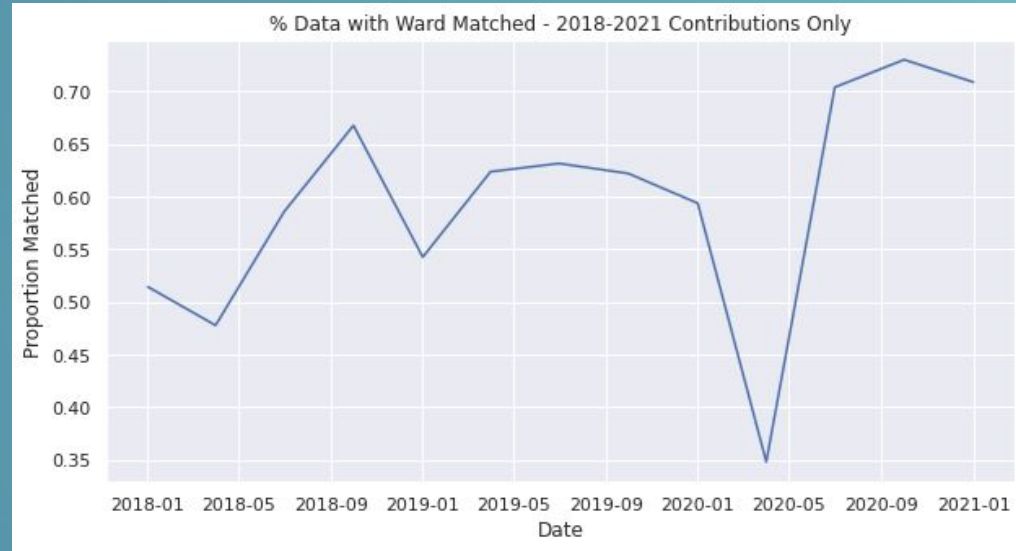
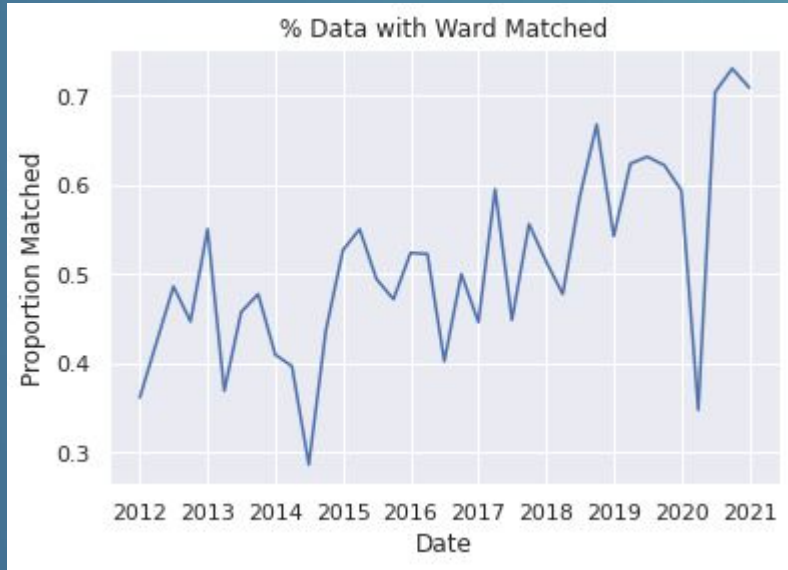
Contributions Data - Trends & Major Recipients



- Almost all lobbyist contributions are for specific aldermen, presumably for funding re-election campaigns
- Spikes in 2015 and 2019 indicate large one-off donations

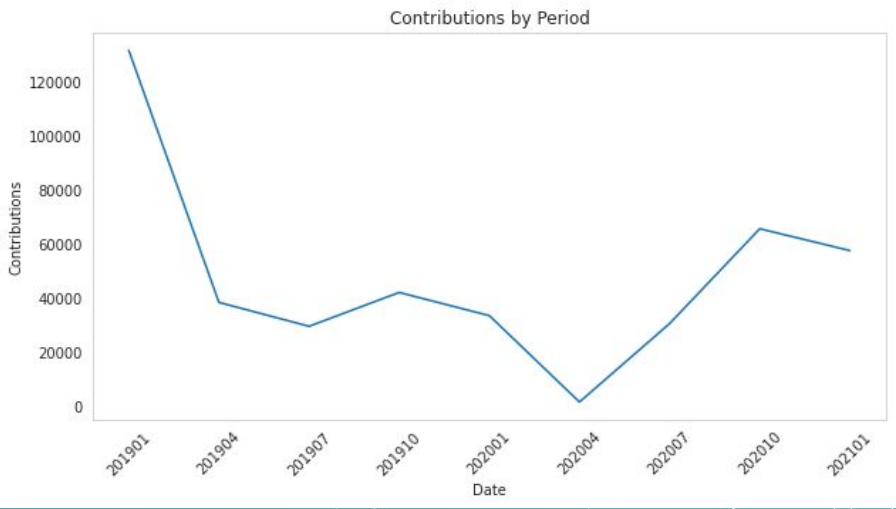
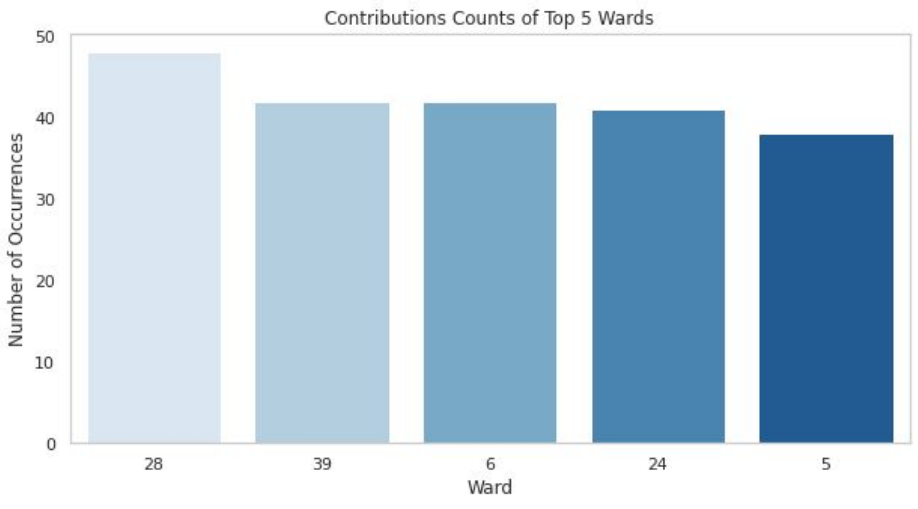


Contributions Data - Success Rates for Matching Recipient Titles with Aldermen/Ward Titles



- Recipient names from the lobbyist contribution data were matched against current aldermen in Chicago. The turnover rate for aldermen is low, but it does make sense for data from 2012 contributions to be less likely to match an existing alderman than 2020 data.
- Our recommendation would be to limit the contributions data to 2018-today to ensure better results.

Contributions



Project requirements

- Executive Summary
- Problem Statement/Research objective(s)
- Exploratory Data Analysis
 - Data analysis, visualizations, data mining techniques
- Data Preparation/Feature Engineering
 - Handling of features, assumptions, and tests
- Methodology and various tools used in the process
 - At-least 4 ML/DL Models implemented and evaluated
- Insights
 - Scope for improvement
 - Assumptions
 - Feature Engineering
 - Data Preparation
 - Model Metrics
- Lessons Learned
 - scope for improvement
- Recommendations
 - Next Steps
 - Methods that can be used
 - Datasets that can add value to the existing analysis
- References
 - Literature review, URLs



Requirements

- **Executive Summary**
- **Problem Statement/Research objective(s)**
- Exploratory Data Analysis
 - **Data analysis, visualizations, data mining techniques**
 - Clarity in underlying assumptions (if any), unit of analysis, feature interactions.
- Data Preparation/Feature Engineering
 - **Handling of features, feature extraction/engineering**
 - **Variable transformations/data scaling**, assumptions, and tests
- Methodology and various tools used in the process
 - **Model selection, descriptions, evaluation approach and key decisions**
 - **At-least 4 ML/DL Models implemented and evaluated**
 - Model deployment strategy (automation)
- Findings and Conclusions
 - **Model Results, performance results, visualizations**
 - **Validating assumptions and impact (\$/hrs.) based on the problem statement**
 - Practicality for the business use and any possible extension to other areas.
- **Lessons Learned** and Recommendations
 - Next Steps along with additional methods/algorithms/models that can be used
 - **Third party datasets that can add value to the existing analysis**
- References
 - **Literature review, URLs**