GROUP 9 EDA REPORT

Kyle Chang, Junhan Yu, Jared Coffey

3/25/2023

1. SUMMARY/OVERVIEW

For our STAT 203 final project, we chose to cover the relationship between gun violence and food insecurity in Rochester. To do this, we gathered data from the RPD Open Data Portal of every shooting incident recorded from 2000-2023 and data from FoodLink indicating the food insecurity rates of every NYS zip code (see section 5 for main findings).

1.1 Initial Findings

Our initial hypothesis consisted of two main parts: "Gun Violence Rates in Rochester are positively correlated with the Food Insecurity Rates in Rochester" and "Given that a correlation exists between the rates of Gun Violence and Food Insecurity in the City of Rochester, then this statistic will also indicate other demographically-based factors".

2. DATA PRE-PROCESSING

2.1 DATA CLEANING

When looking through our datasets, we noticed how unnecessarily large they were: the RPD ODP records [3] included gun violence incidents dating back to 2000 and FoodLink calculated the food insecurity rates of every zip code in New York State. As you could imagine, this amount of data would both slow down our EDA process and cloud our focus on Rochester.

As for our Food Insecurity Data, we had two sets: our data obtained from FoodLink.org[1] (our original dataset from when we wrote our proposal), and the more recent data that we obtained from FeedingAmerica.org [2]. We found that since the latter dataset was a national report, it had very little data that was specific to Rochester and the city's Zip Codes. Thus, we decided only to use the FoodLink Data, as it was just focused on the City of Rochester's Food Insecurity Levels.

Both datasets included variables unrelated to the relationship in question. So, we decided to clean such excessive data. Now, our gun violence data spans from 2022 to 2023 and our FoodLink data only covers Rochester zip codes.

GUN VIOLENCE DATA AFTER DATA CLEANING							
	X	Y	Address	Occurred_Date	Occurred_Year	Latitude	Longitude
0	-77.610890	43.184163	442 Remington St	2022/06/01 04:00:00+00	2022	43.184163	-77.610890
1	-77.598893	43.181793	904 Hudson Ave	2022/05/29 04:00:00+00	2022	43.181793	-77.598893

FOOD INSECURITY DATA AFTER DATA CLEANING				
	Zip_Code	Latitude	Longitude	Food Insecurity
124	14626	43.2141	-77.7135	0.094
127	14623	43.0881	-77.6425	0.167

2.2 DATA DISCRETIZATION

Continuing on, we encountered our first problem: unlike the food insecurity dataset, our gun violence dataset did not record the zip codes of each incident. However, we were given the latitudes and longitudes, which we converted to zip codes using the Nominatim package. This made it possible to categorize our data, grouping them by zip codes, and recording how many gun violence incidents were recorded within each zip code. This will be useful not only for merging our datasets together, but also for visualizing it (which we cover in the next section).

2.3 DATA MERGING

Now that both datasets are organized by zip code, we can merge the two datasets. We will also be using min-max normalization on our two attributes. View a sample of our dataset below.

Scaled Dataset Using Min-Max Normalization						
	Zip_Code	Latitude	Longitude	Food Insecurity	Number of Shootings	Gun Violence
0	14626	43.2141	-77.7135	0.063063	0	0.0
1	14623	43.0881	-77.6425	0.282282	0	0.0

3. EXPLORATORY DATA ANALYSIS

3.1 PLOTTING OUR DATA

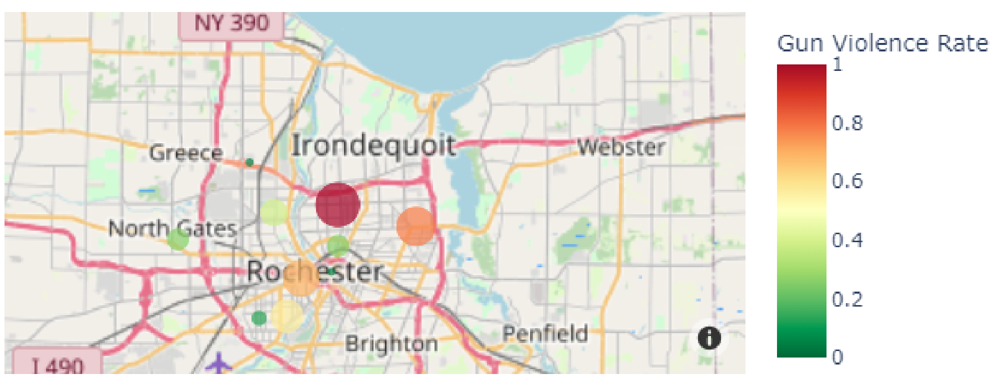
Visualizing data is a key to better understanding it, so that’s what we’ll do with our now fully merged dataset. Before plotting, we quickly obtained a quick data summary of our newly integrated and merged dataset:

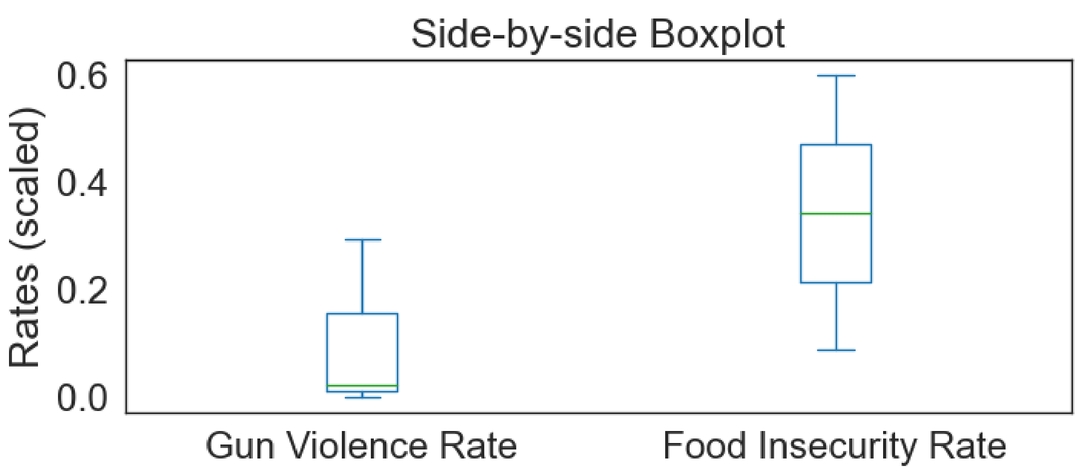
DATA SUMMARY

	Food Insecurity	Gun Violence
count	20.000000	20.000000
mean	0.372222	0.208696
std	0.306453	0.312445
min	0.000000	0.000000
25%	0.087838	0.000000
50%	0.342342	0.021739
75%	0.597598	0.293478
max	1.000000	1.000000

To the average person, this summary may provide further confusion. Fortunately, there exists a couple of sure-fire ways to visualize this in an understandable manner: scatter maps and box plots. Because we will be using zip codes as an identifier, we can plot them onto a map of Rochester (see below).

Rochester Zip Codes by Gun Violence Counts





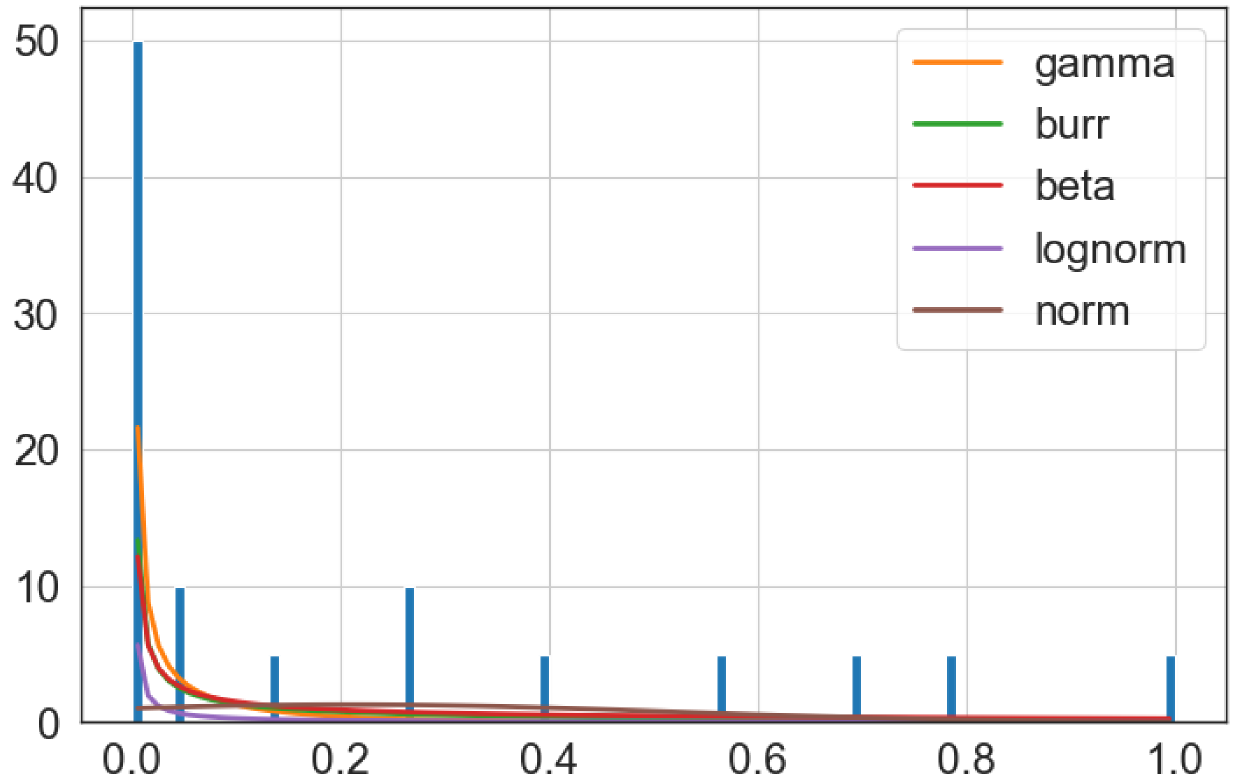
As shown above, our food insecurity data seems to be symmetric given that the the 25th, 50th, and 75th percentiles are evenly spaced between each other. This is further supported by the summary data, which shows that the mean and median only differ by .03.

On the other hand, the same could not be said for the gun violence data. In this case, the distance between the 50th and 75th percentiles is significantly greater than the distance between the 25th and 50th percentiles. In the summary data shown in the previous section, the mean is also much larger than the median for gun violence. All of these indicate a strong positive skew. This means that our gun violence data will have many outliers in the higher ranges of shootings rate.

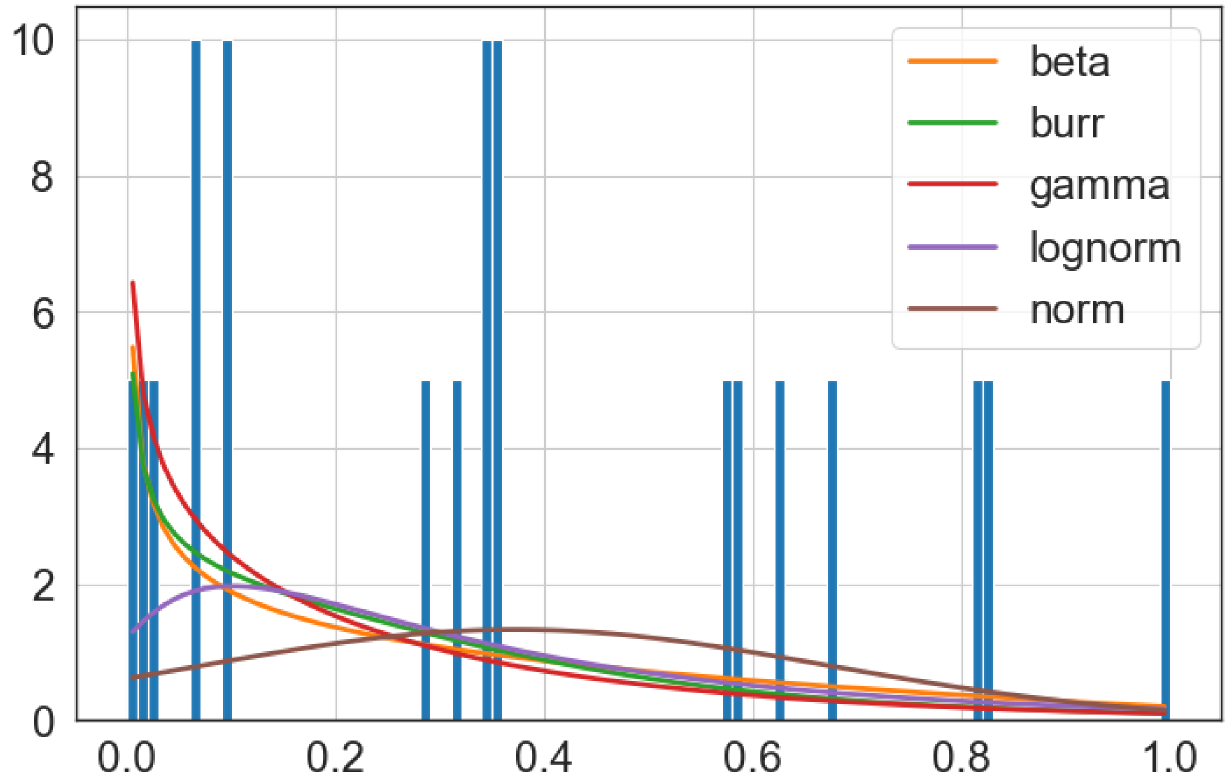
We can thus conclude that - while higher levels of food insecurity seem to be spaced evenly (and centered around the southern part of the city), the levels of gun violence seem to be more sporadic.

4. FITTING A DISTRIBUTION TO OUR DATA

Another important part of this process is fitting a distribution to our data. This will allow us to better understand the way that our variables may be distributed, giving us a bigger picture of how they will behave given a larger sample size.
Here, we can see five possible distributions and how they look when plotted against our gun violence data. Using a function to return the best of these five, we get that this data can be fitted into the gamma distribution.



Now, we will do the same process for our food insecurity data, and discover that the beta distribution best fits this set (see below).



In order to compare gun violence rates to that of food insecurity rates, we had to derive our own metric, as there was no concrete measure for gun violence within our dataset. In order to keep things consistent however, we performed min-max normalization on both the Food Insecurity RateColumn and Gun Violence Rate Column. Our method consisted of

$$Rate(ZIP\ Code) = \frac{corresponding\ value(ZIP\ Code) - (min\ rate\ of\ all\ ZIP\ Codes)}{(max\ rate\ of\ all\ ZIP\ Codes) - (min\ rate\ of\ all\ ZIP\ Codes)}$$

While we realize that this is not 100% accurate nor is it the best way of doing it, we figured that this derived metric is good enough, since we are also making other assumptions. With more time, we would hope to use other dividing metrics like using addresses, streets, or polygonal data in order to compare and contrast our results obtained from using ZIP Codes.

5. Summary of Final Findings

We found that while there *is* a positive correlation between these two variables, the coefficient 0.65 is not overly high, and thus it seems that other factors are at play that contribute to each problem. This signaled that not only is there not provable causation between the two factors, there is also not too much correlation (which we had originally hypothesized and counted on for our second hypothesis).

After taking a look at both of the variables' distributions, our group concluded that while our food insecurity data was very well modeled by a gamma distribution, our gun violence data was all over the place. In fact, our gun violence data does not seem to fit any distribution particularly well at all. As discussed above, the results from 2022 are very sporadic, with clusters of data sprinkled throughout different parts of the city. Our group theorized that this is mainly what caused the relatively mid-level correlation between Gun Violence and Food Insecurity.

Our original plan consisted of testing whether or not our correlation would provide a good indicator of other demographically based variables such as race or age demographic (i.e.: Does an area with both high food insecurity and high gun violence have any trends that we can extract?). While we had originally planned to do this with a Time Series Regression model, we felt that our correlation was not statistically significant enough for a regression analysis to tell us anything new, let alone be reliable in any sense. Given more time perhaps, we could do a deep-dive into each of these other factors given to us and how they relate to our findings, but we found that they could each inhabit an entire report in and of themselves.

6. APPENDIX

Gun Violence Data Dictionary

Attribute	Data type	Description
X	Continuous	Longitude of the shooting
Y	Continuous	Latitude of the shooting
ID	Nominal	ID of the event
Case_Number	Nominal	Case number assigned
Address	Nominal	The address at which the event occurred
Occurred_Date	Ordinal	Date of shooting

Occurred_Month	Ordinal	Month of shooting
Occurred_Year	Ordinal	Year of shooting
Crime_Type	Nominal	Type of crime (Shooting or Homicide)
Multiple_Shooting	Nominal	Multiple victims
Gender	Nominal	The gender of the shooter
Race	Nominal	The race of the shooter
Ethnicity	Nominal	The ethnicity of the shooter (Hispanic or Non Hispanic)
Victim_Age	Discrete	The age of victim
Victim_Age_Band	Nominal	The age group of victim
Latitude	Continuous	Latitude of the shooting
Longitude	Continuous	Longitude of the shooting
ObjectID	Discrete	ID of the object in the dataset

Food Insecurity Data Dictionary

Attribute	**Data type**	**Description**
Zip Code	Nominal	The zip code whose FI is measured
Latitude	Continuous	The latitude of the area measured
Longitude	Continuous	The longitude of the area measured
Food Insecurity	Int, Continuous	The FI rate of the corresponding area

Map the Meal Gap National Report*

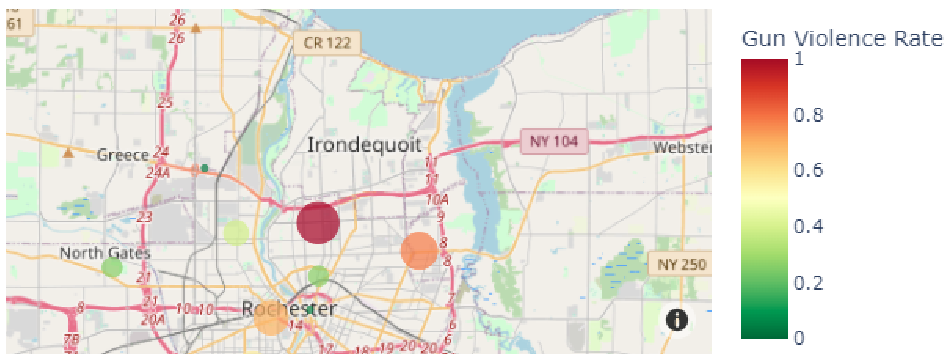
Attribute	Data type	Description
FIPS	Nominal	Federal Information Processing Standard - unique identifier of a county
State	Nominal	The state of the county
County, State	Nominal	The county and state of the area described
Year	Ordinal	Year of the data provided
Overall Food Insecurity Rate (1 Year)	Continuous	Overall FI rate in the given year and county
# of Food Insecure Persons Overall (1 Year)	Discrete	Total number of food insecure persons in the given year and county
Food Insecurity Rate among Black Persons (all ethnicities)	Continuous	FI rate of Black citizens in the area
Food Insecurity Rate among Hispanic Persons (any race)	Continuous	FI rate of Hispanic citizens in the area
Food Insecurity Rate among White, non-Hispanic Persons	Continuous	FI rate of White, non-Hispanic citizens in the area
Child Food Insecurity Rate (1 year)	Continuous	FI rate of children in the area
# of Food Insecure Children (1 year)	Discrete	Total number of food insecure children
% food insecure children in HH w/ HH incomes below 185 FPL	Continuous	The percentage of food insecure children that live in households with an income below 185% of the Federal Poverty Level
% food insecure children in HH w/ HH incomes	Continuous	The percentage of food insecure children that live in households with an income above 185% of the Federal Poverty Level

above 185 FPL		

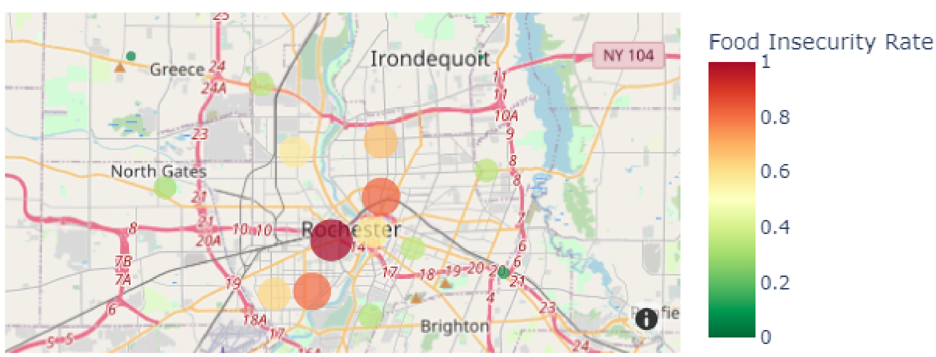
** note: this data ended up not being used for reasons stated above*

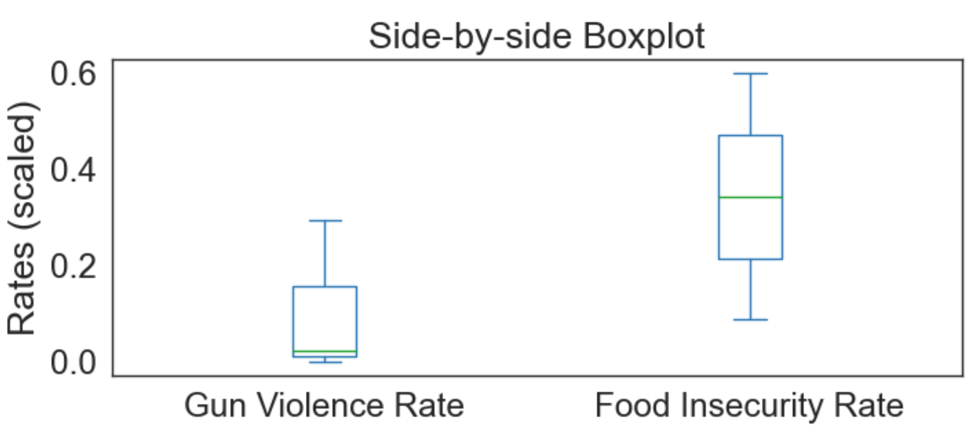
Plots and Other Visualizations

Rochester Zip Codes by Gun Violence Counts

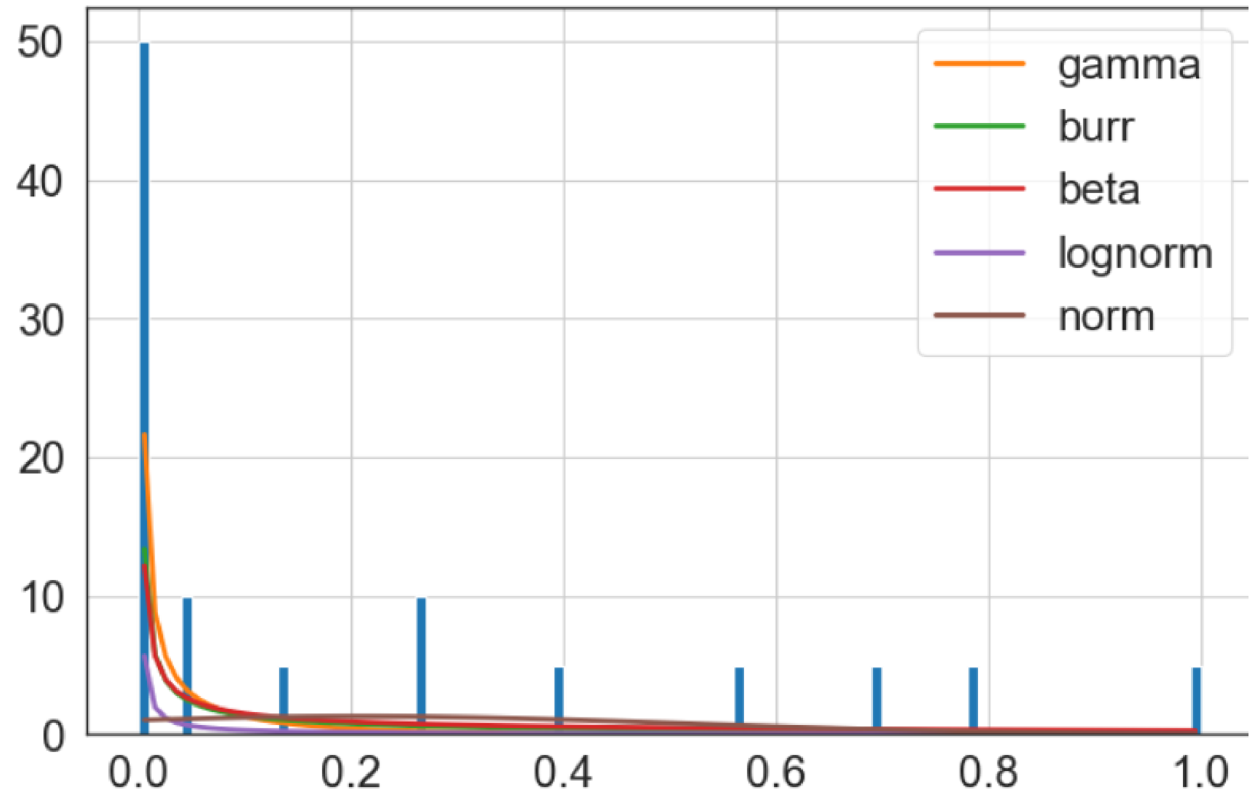


Rochester Zip Codes by Food Insecurity Rate

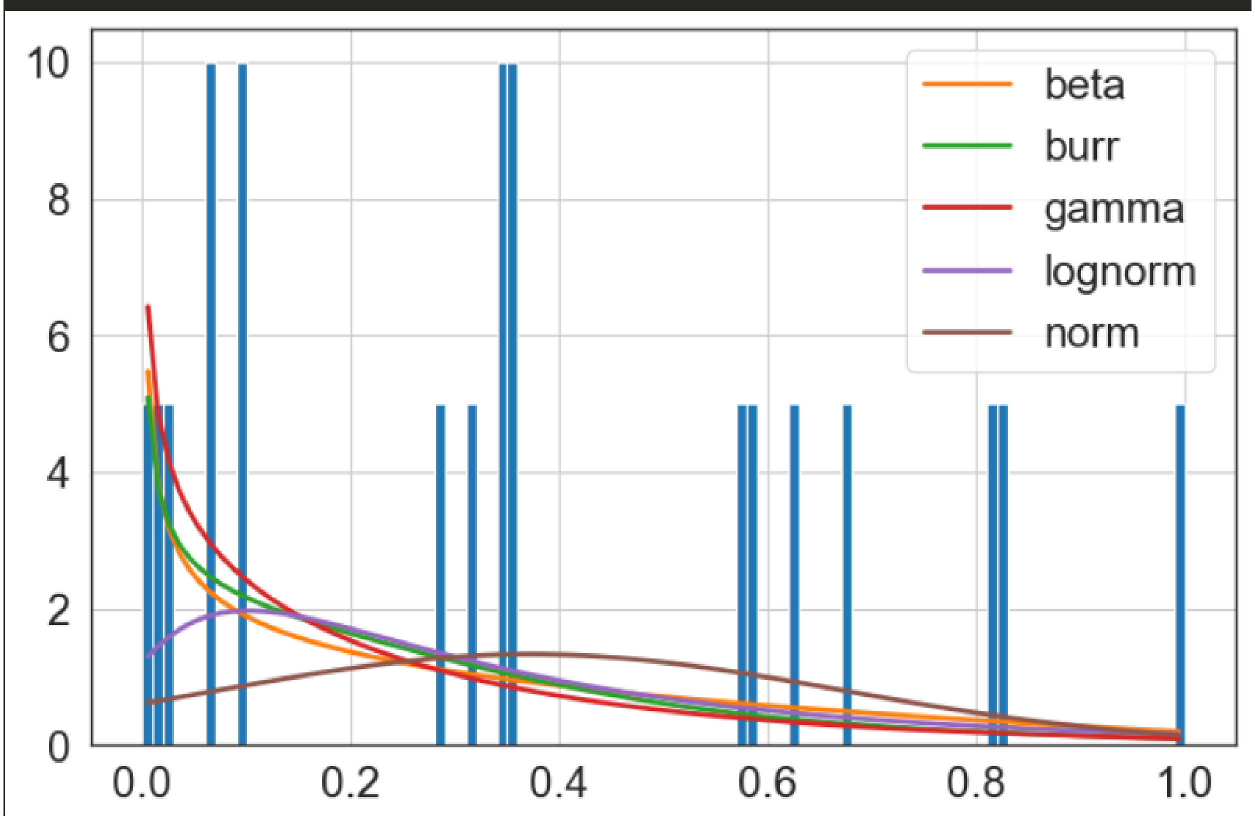


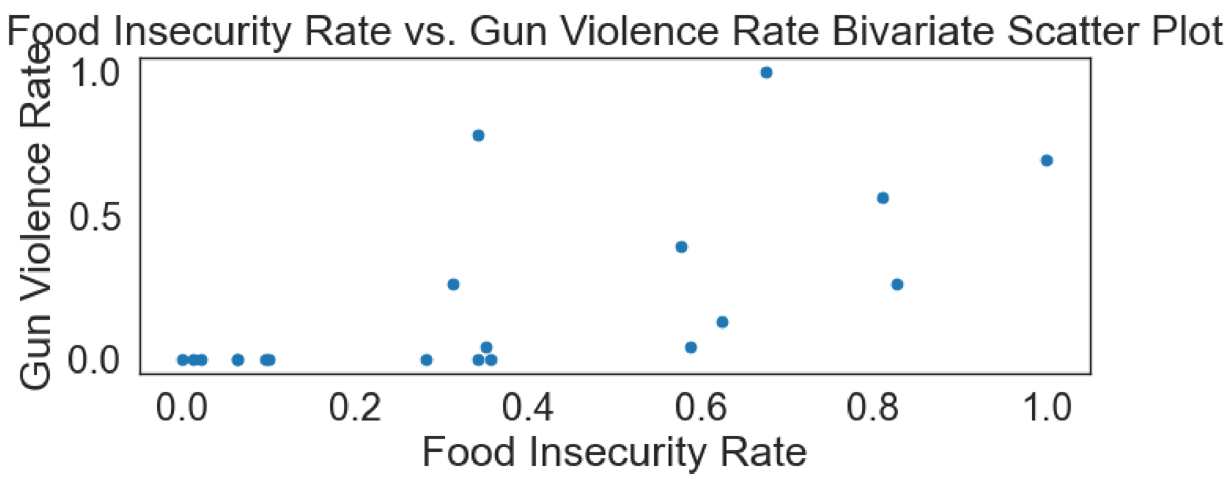


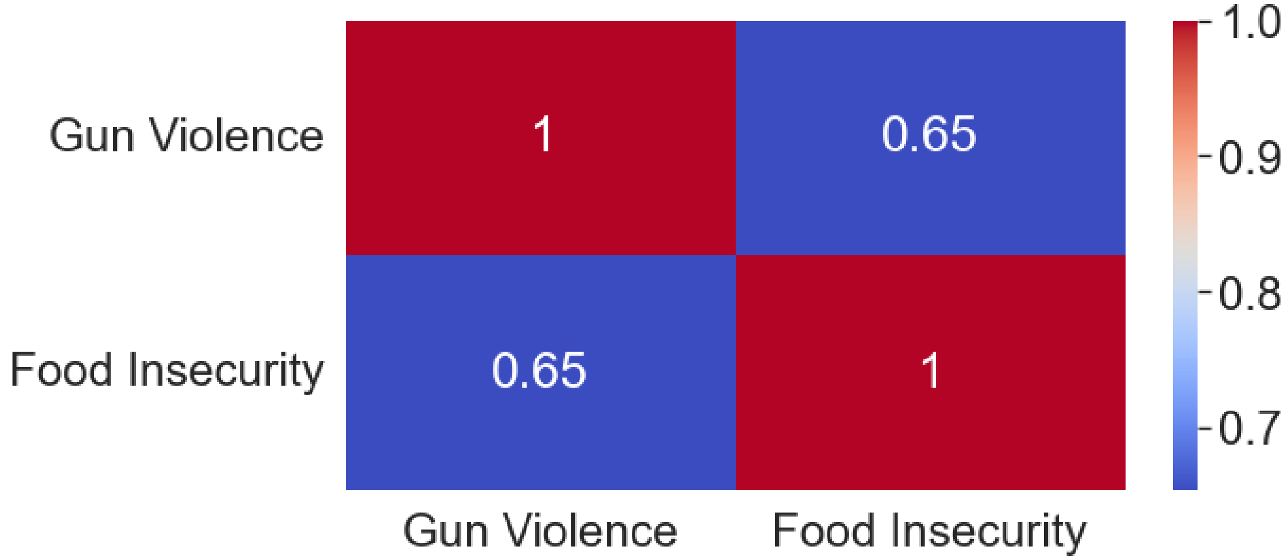
FITTING DISTRIBUTION TO GUN VIOLENCE RATE



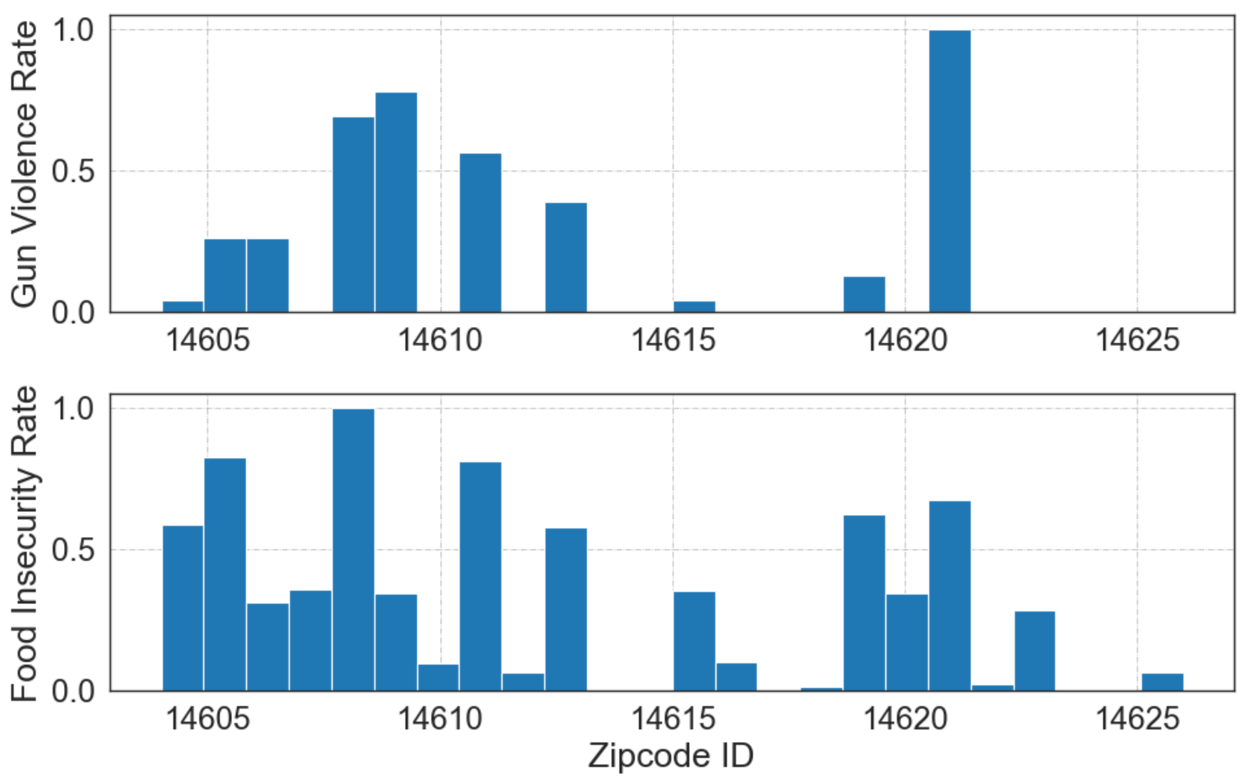
FITTING DISTRIBUTION TO FOOD INSECURITY RATE




Food Insecurity Rate vs. Gun Violence Rate Bivariate Scatter Plot
Gun Violence Rate
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Food Insecurity Rate


Gun Violence
Food Insecurity
1
0.65
0.65
1
Gun Violence
Food Insecurity
1.0
0.9
0.8
0.7

Relative Frequency of Gun Violence Incidents per Zip Code



Dataset Sourcing:

[1] Dwyer, M. (2018), "New food insecurity data show level of need in Rochester, other communities," *Foodlink Inc*, Available athttps://foodlinkny.org/new-food-insecurity-data-show-level-of-need-in-rochester-other-communities/.

[2] "Overall (all ages) Hunger & Poverty in the United States | Map the Meal Gap" (n.d.). Available athttps://map.feedingamerica.org.

[3] "Rochester NY Shooting Victims | Rochester NY Shooting Victims | Rochester, NY Police Department Open Data Portal" (n.d.). Available athttps://data-rpdny.opendata.arcgis.com/datasets/rochester-ny-shooting-victims/explore?location=43.180005%2C-77.596549%2C5.00.