Angela Baker
January 2019

STI Cases in the Chicago area, a study from 00-14

This data had many challenges. First and foremost it is a small data set with only 14 years of data collected from 88 key community areas in Chicago. Unable to visualize all the data and looking for the most complete sets to compare I settled on the areas which had over 80,000 people per community area as they were likely to have the best insight. Into how the population as a whole might be predicted. I chose not to compare the lesser populated areas as many areas had missing data and therefore a complete analysis could not be completed. This dataset(s) was obtained from: https://www.healthdata.gov/search/type/dataset?query=STI&sort_by=changed&sort_order=DESC and was cleaned and wrangled using a multitude of techniques.
First the data was wrangled via Excel, as it is not as foreign to me, however after the visualization module I then went back and cleaned it using Python techniques with pandas and matplotlib.
Represented below are the basic trends seen in Lake View, Near North Side, West Town, and Austin, for both Males and Females in both Chlamydia and Gonorrhea studies.

Null: That the spread across the Chicago area communities can be represented as a whole and not by community area boundaries.

Alternative: That the Chicago area communities differ in a statically significant manor.

Visualizations:

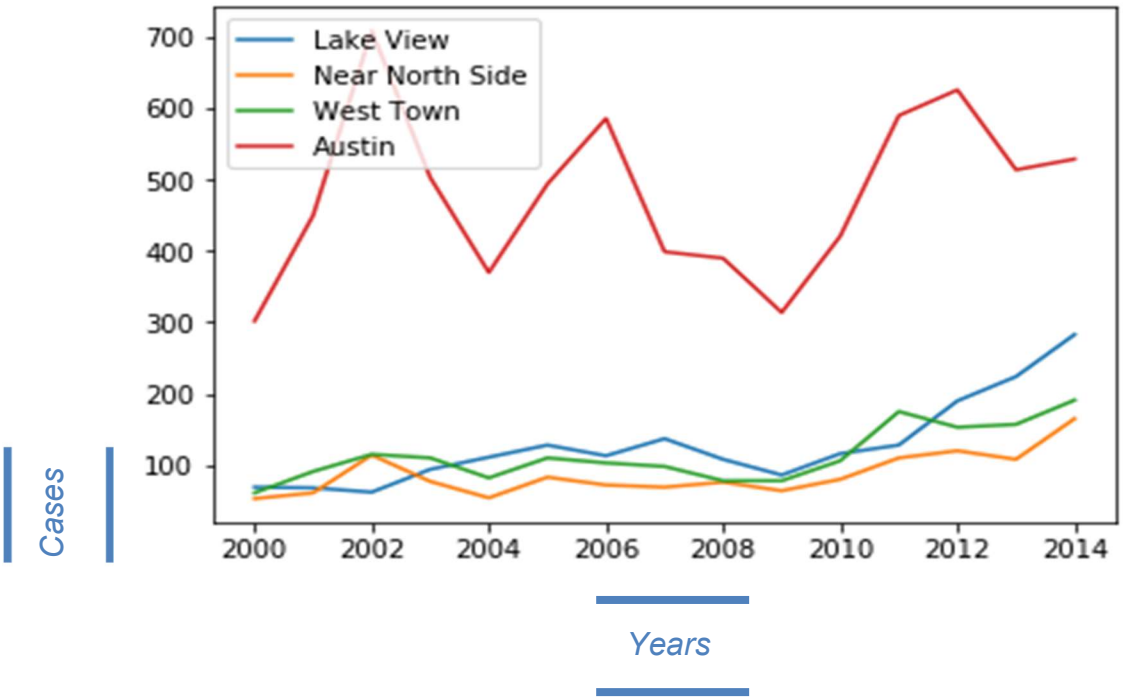


Figure 1. Chlamydia Cases for Males

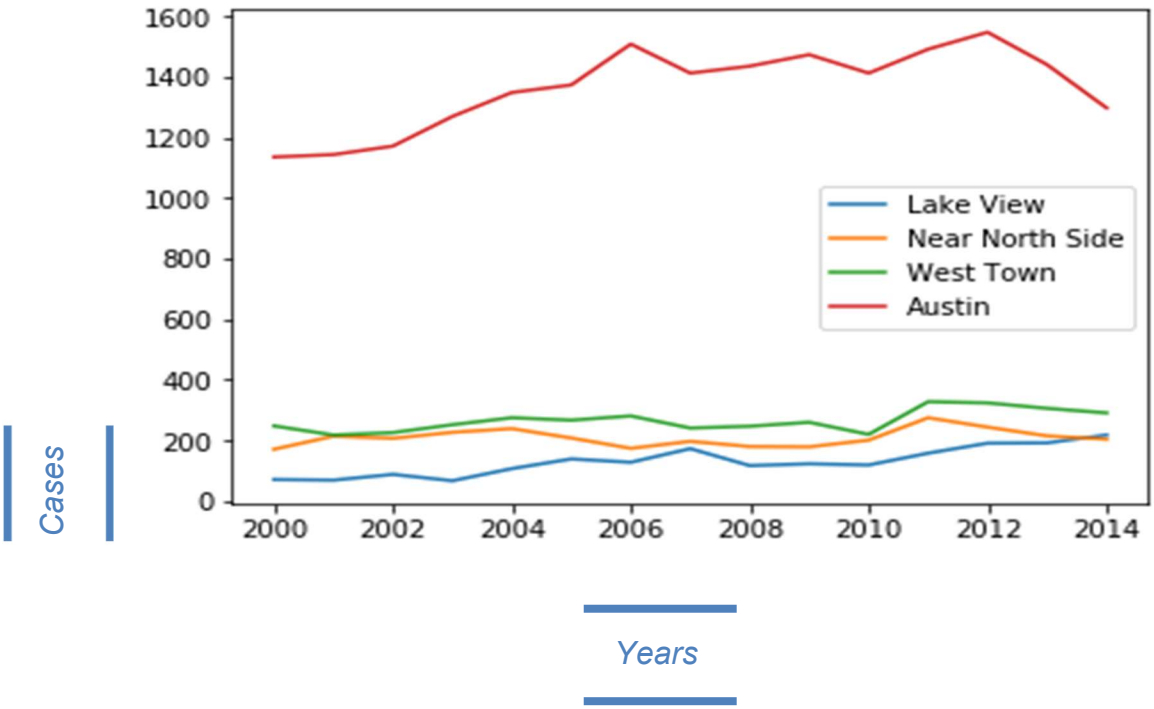


Figure 2. Chlamydia cases for Females

Table 1. Chlamydia Means- Mean amount of cases for four highest populations in the Chi-town area

	F	M
Lake View	130.6667	127.8
Near North Side	209	87.06667
West Town	265.6	184.6667
Austin	1364.6	113.8667

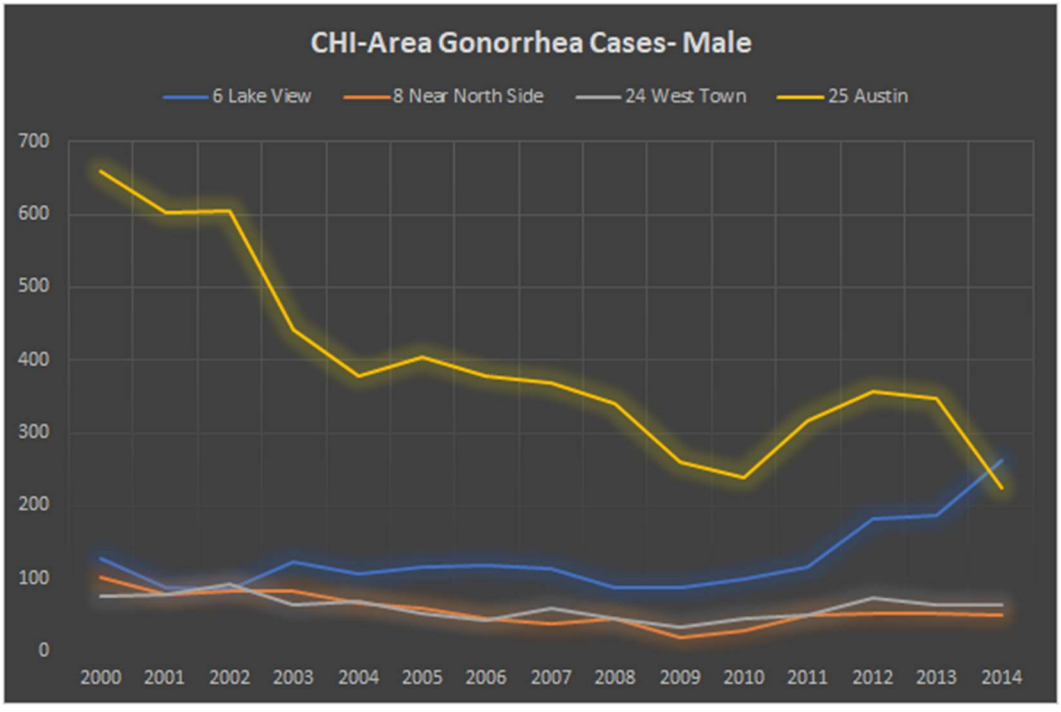


Figure 3. Austin's rate comes down within the other ranges

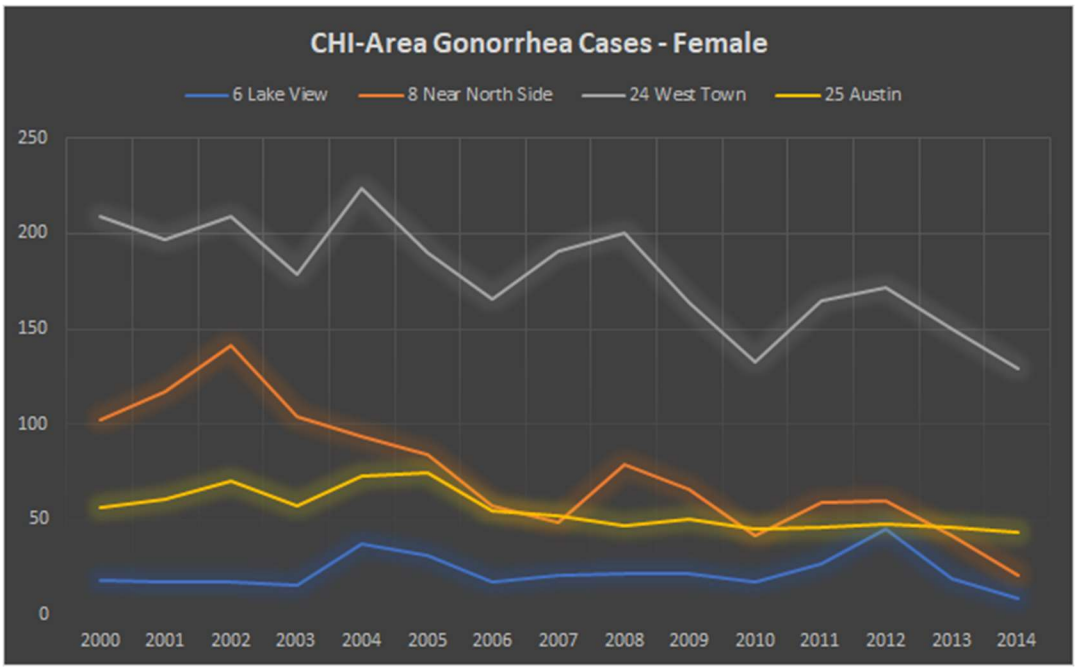


Figure 4.

Statistics:

	Gonorrhea		Gonorrhea-BS (Bootstrapped resampled)	
	F	M	F	M
Lake View	22.33333	126.4667	23.06667	132.3333
Near North Side	74.46667	56.8	82.73333	57.13333
West Town	178.5333	60.33333	162.9333	64.53333
Austin	54.93333	395.0667	53.2	406.4667

Table 2. Original data and resampled bootstrap data- 1 sample

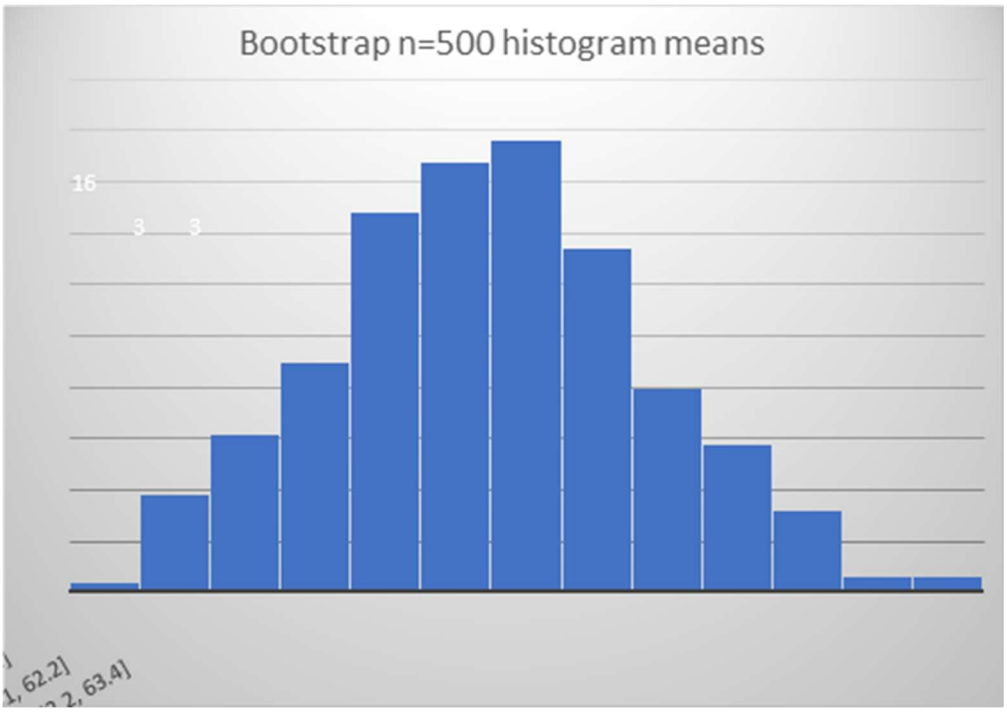


Figure 5. Sample of Austin data gon-Female n=500 The distribution is normal.

As we can see this data does not have any outliers and is clearly within the possibility of reason. However, Austin as a whole seems to be an outlier as it towers every other Community areas data, both even those with similar populations and land area as well as access to clinics.

mean	54.95422488
confidence interval	1.96
stdev	2.705598805
n	501
margin of error	0.236919391
lower bound	54.71730549
upper bound	55.19114427

Table 3. Stats of the n=501 sample Gon-Female

Looking at the original data for Gon-Female for Austin and West Town (the two largest in population), running a regression we find the p-value of 0.002173 which is < 5%, so in this case we can throw out the null hypothesis that the areas are representative of the whole population. We can see just from this data that even the two largest in populations are statistically different from one another and thus each community area must be treated differently and likely has other contributing factors.

Overview:

What the data says is that all the samples can be considered as independent and statically different. However, what a client will be looking for is what happened in these cases to create such a wide spread and such different cases numbers for areas that appear to have relatively the same size of populations.

Digging Deeper:

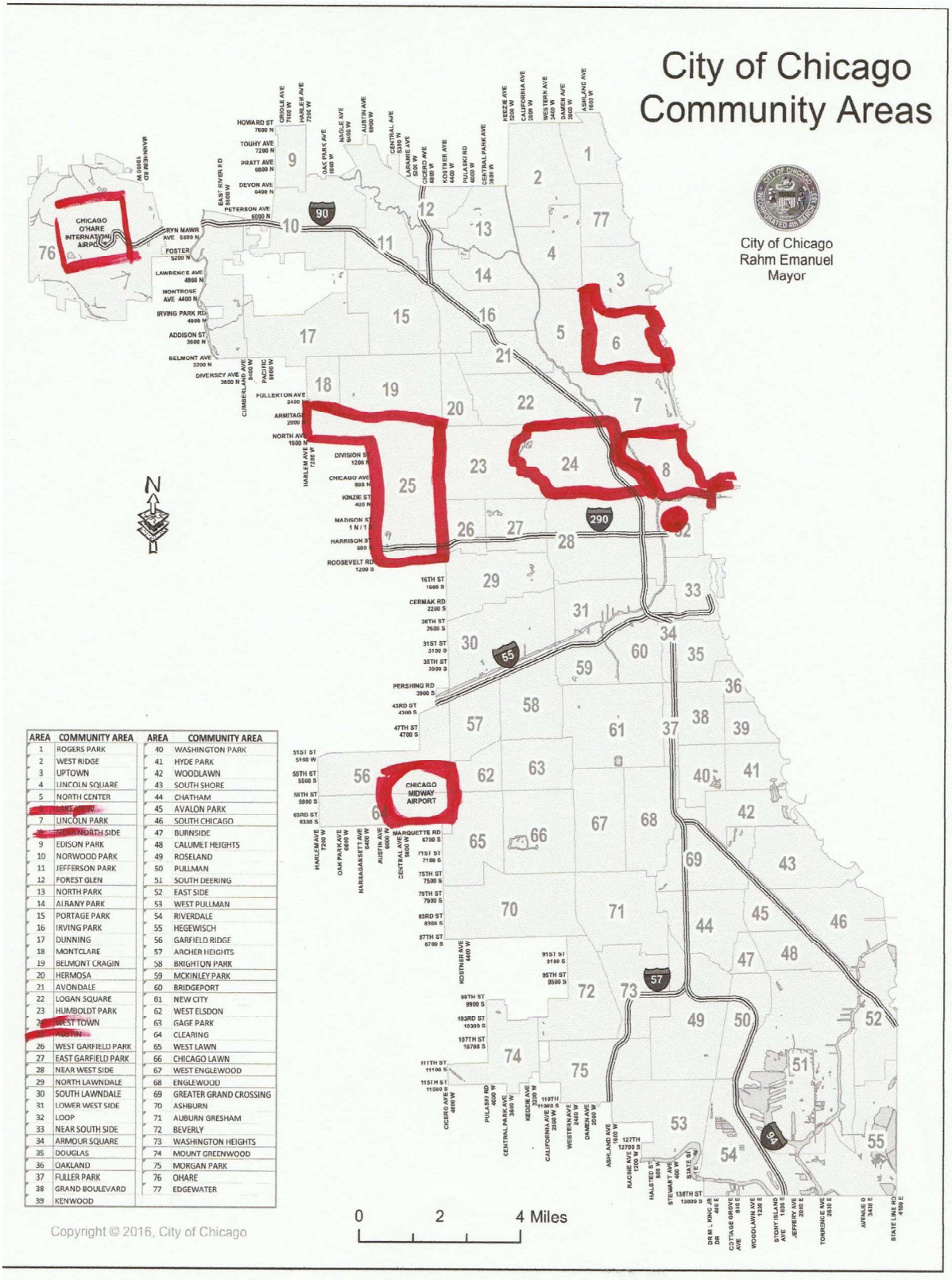
Below is a map of the Chicago area Community Neighborhoods
6- Lake View, 8- Near North Side, 24- West Town, 25- Austin The 4 most populated neighborhoods. To the upper left and mid left we see where the two major airports are and the large dot in community section 32 designates the Downtown area.

By cross-referencing Zip Codes and Community areas via http://chicago-zone.blogspot.com/2014/03/chicago-zip-code-map-locate-chicago.html
And Free Clinics in the Chicago area :
https://www.cityofchicago.org/dam/city/depts/cdph/policy_planning/PP_Web%20Health%20Care%20Facilities%20by%20Region.pdf

Area	Population_2010	Clinics Available
Lake View (6)	94,368	3
Near North Side (8)	80,484	2
West Town (24)	82,236	5
Austin (25)	98,514	7
Lower West Side (31)	35,769	8
New City (61)	44,377	8

South Lawnsdale(30)	79,288	7

Table 4. Population v. free clinics available using both the highest populations [Lake View, Near North Side, West Town, Austin] and Communities with the largest number of available clinics [Austin, Lower West Side, New City]



AREA	COMMUNITY AREA	AREA	COMMUNITY AREA
1	ROGERS PARK	40	WASHINGTON PARK
2	WEST RIDGE	41	HYDE PARK
3	UPTOWN	42	WOODLAWN
4	LINCOLN SQUARE	43	SOUTH SHORE
5	NORTH CENTER	44	CHATHAM
6	LAKE VIEW	45	AVALON PARK
7	LINCOLN PARK	46	SOUTH CHICAGO
8	NEAR NORTH SIDE	47	BURNSIDE
9	EDISON PARK	48	CALUMET HEIGHTS
10	NORWOOD PARK	49	ROSELAND
11	JEFFERSON PARK	50	PULLMAN
12	FOREST GLEN	51	SOUTH DEERING
13	NORTH PARK	52	EAST SIDE
14	ALBANY PARK	53	WEST PULLMAN
15	PORTAGE PARK	54	RIVERDALE
16	IRVING PARK	55	HEGEWISCH
17	DUNNING	56	GARFIELD RIDGE
18	MONTCLARE	57	ARCHER HEIGHTS
19	BELMONT CRAGIN	58	BRIGHTON PARK
20	HERMOSA	59	MCKINLEY PARK
21	AVONDALE	60	BRIDGEPORT
22	LOGAN SQUARE	61	NEW CITY
23	HUMBOLDT PARK	62	WEST ELSDON
24	WEST TOWN	63	GAGE PARK
25	AUSTIN	64	CLEARING
26	WEST GARFIELD PARK	65	WEST LAWN
27	EAST GARFIELD PARK	66	CHICAGO LAWN
28	NEAR WEST SIDE	67	WEST ENGLEWOOD
29	NORTH LAWNDALE	68	ENGLEWOOD
30	SOUTH LAWNDALE	69	GREATER GRAND CROSSING
31	LOWER WEST SIDE	70	ASHBURN
32	LOOP	71	AUBURN GRESHAM
33	NEAR SOUTH SIDE	72	BEVERLY
34	ARMOUR SQUARE	73	WASHINGTON HEIGHTS
35	DOUGLAS	74	MOUNT GREENWOOD
36	OAKLAND	75	MORGAN PARK
37	FULLER PARK	76	OHARE
38	GRAND BOULEVARD	77	EDGEWATER
39	KENWOOD		

Figure 6. Map of the Chicago Neighborhoods

As it is glaringly obvious the areas that have the highest need are severely underserved. Austin has the most land area out of the 4 highest populations and is similarly represented as South Lawnsdale and they are on the Western side of the suburbs from the city meaning that they are very closely represented as far as land mass to population ratios. However Lake View, Near North Side, and West Town are all significantly smaller with similar population sizes, meaning that the areas are more densely populated and have significantly less access to free clinics to be tested.

The data earlier showed that Lake View, Near North Side, and West Town were very similar in number of cases across both STI's and gender. Let's now take a look at Austin and South Lawnsdale as their populations, land mass, and clinic availability are similar, lets see if their STI are similar as well.

Looking at Figure 6 below we see that indeed like the rest of the community areas Austin remains the outlyer and as to why is has so many cases the vary so rapidly from year to year. In order to be fully informed there would need to be polling done as to the months that these cases are seen or if there is a particular spike after large events or if the people of Austin just get tested more regularly than other community areas.

At this point in time it would be hard to give any definite answers without speaking with community members and those that also work in the clinics for more insight as to these numbers or perhaps the numbers for Austin were reported inappropriately, that is also an option.

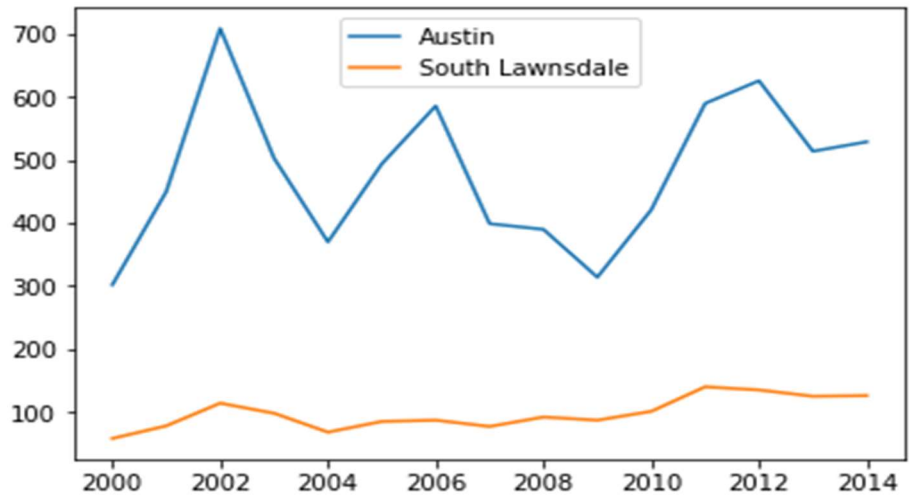


Figure 6. Austin and South Lawnsdale Chlamydia Male cases

Machine Learning:

At this juncture running a PCA on the Austin South Lawnsdale data will give an idea on how varied the data are:

```
array([[-181.8670012 ,   11.94033715],
       [ -32.55712911,   15.19027502],
       [ 228.89059846,   19.88982275],
       [  22.9076067 ,    3.67232895],
       [-113.1391264 ,   12.63259762],
       [  11.99614651,   15.11523227],
       [ 103.18865499,   27.4410749 ],
       [ -83.09260105,    8.25009665],
       [ -89.65142176,   -7.96661913],
       [-165.50478784,  -14.84170859],
       [ -57.62920934,  -12.03821729],
       [ 115.37895388,  -24.29282707],
       [ 150.16406426,  -13.75735861],
       [  37.9710737 ,  -21.28948112],
       [  52.94417821,  -19.94555351]])
```

As expected the variance between the data is very wide and therefore in this set a method would be easily trainable to categorize any new data as Austin or South Lawnsdale.

More helpful data however would be if these cases were broken down by month and then created a model which predicted the number of cases by month or then collectively by year, or even if there was full data on the entire set (meaning that all the community areas actually reported (there are only a few full sets in this data) then there could be a heat map produced of various years and then those years could be compared.

Additionally, this data set is not set up very well for machine learning as it is so small, there would not be many useful insights to draw from creating a model to run classification data or even unsupervised learning as the set is unable to supply enough information for learning analysis. If this set were broken down by month it would be worth building a model with predictive capabilities for month to month based on past behavior and training on those test sets for a future set of predictive abilities.

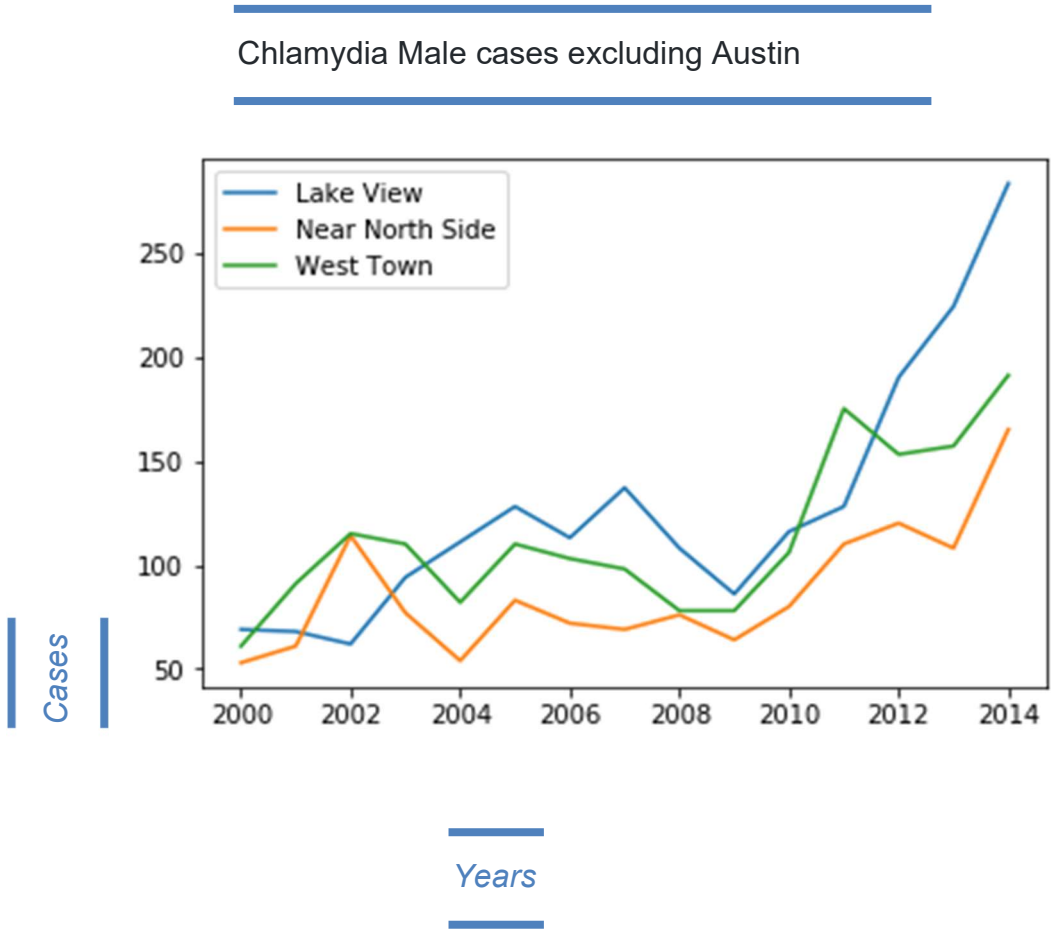


Figure 7. High Population without Austin

Here we can more clearly see the rise in STI despite the fact that in these areas the population change has been negative from 2000-2010, however missing from this data is how the population has changed from 2010 till now or even the end of the data frame as there is no other census as it has not been 10 years.

Conclusion:

This data set was extremely small and because of that had some pretty big limitations. Additionally most of the data was just better visualized and not run through statistical modeling or predictive machine learning techniques. I now understand when I first got this data that my mentor had said it will likely be problematic. Cleaning this data was a challenge as everything was labeled differently additionally only having 14 years and not enough data points.
In dissecting these data sets I have learned a lot of what raw files should have and what they shouldn't. As I grow into my big data shoes and get more data analysis tools I know it will become easier and I will find more innovative ways to solves complex issues

and rough data. This capstone was filled with many ups and downs but I have learned a lot in the many iterations that I have had of this capstone, running and rerunning models, becoming close friends with seaborn, pandas, and python coding in general. I look forward to the next challenges and where a bigger more robust data set can take me.