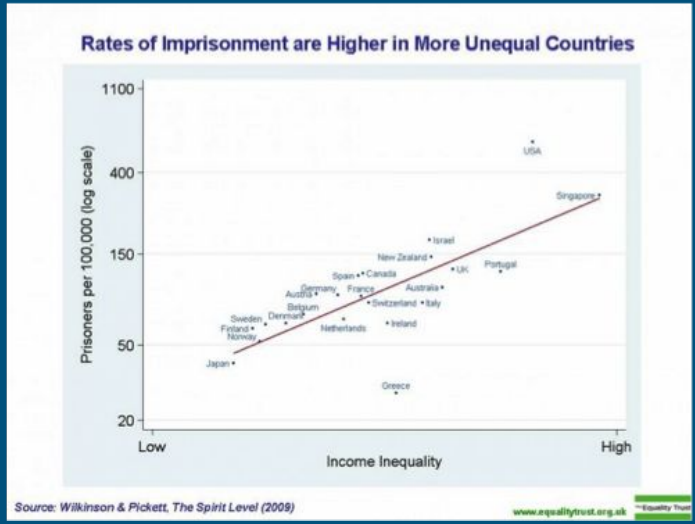
Analysis of Violent Crime

By Poisson Criminals
Tuc Baldecchi, Ayush Dev, Ujvala Cheedella, and John Luksas

Problem Description

- Violent Crimes in the US are a complex and difficult topic to comprehend
- Income inequality and the correlation between crime rates is a controversial topic
- We will look at the city of Chicago to test this correlation
    - The most popular dataset in US National Crime Dataset
    - Most highly rated dataset



Description of Data

Primary Data: Primary data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system from 2001 to 2018.

Secondary Data: Secondary data is household income data that is extracted from the Chicago Data Portal where it includes census data regarding the six main socioeconomic factors, that all are utilized to calculate the hardship index.

- We used both of these databases to create a master data set that captured parts of the crime and income data

	Community.Area	Time	Primary.Type	Arrest	Domestic	District	Year	Latitude	Longitude	Community.Area.Name	Per.Capita.Income	Harship.Index
1	1	15	BATTERY	TRUE	TRUE	24	2003	42	-87.7	Rogers Park	23939	39
2	1	22	BATTERY	FALSE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
3	1	9	THEFT	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
4	1	1	ROBBERY	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
5	1	12	THEFT	FALSE	FALSE	24	2004	42	-87.7	Rogers Park	23939	39
6	1	22	NARCOTICS	TRUE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
7	1	14	BATTERY	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
8	1	22	NARCOTICS	TRUE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
9	1	18	CRIMINAL TRESPASS	TRUE	FALSE	24	2004	42	-87.7	Rogers Park	23939	39
10	1	13	MOTOR VEHICLE THEFT	FALSE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
11	1	19	OTHER OFFENSE	FALSE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
12	1	19	OTHER OFFENSE	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
13	1	3	BATTERY	FALSE	FALSE	24	2015	42	-87.7	Rogers Park	23939	39
14	1	9	BATTERY	FALSE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
15	1	19	BATTERY	TRUE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
16	1	13	CRIMINAL DAMAGE	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
17	1	20	BATTERY	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
18	1	22	BATTERY	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
19	1	9	NARCOTICS	TRUE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
20	1	11	ROBBERY	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
21	1	13	THEFT	FALSE	FALSE	24	2004	42	-87.7	Rogers Park	23939	39
22	1	11	BATTERY	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
23	1	12	THEFT	FALSE	FALSE	24	2003	42	-87.7	Rogers Park	23939	39
24	1	9	ROBBERY	FALSE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
25	1	12	THEFT	FALSE	FALSE	24	2017	NA	NA	Rogers Park	23939	39
26	1	23	BATTERY	TRUE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
27	1	22	CRIM SEXUAL ASSAULT	FALSE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39
28	1	13	CRIMINAL DAMAGE	TRUE	FALSE	24	2004	42	-87.7	Rogers Park	23939	39
29	1	17	NARCOTICS	TRUE	FALSE	24	2015	42	-87.7	Rogers Park	23939	39
30	1	12	THEFT	FALSE	FALSE	24	2002	42	-87.7	Rogers Park	23939	39

Important Variables

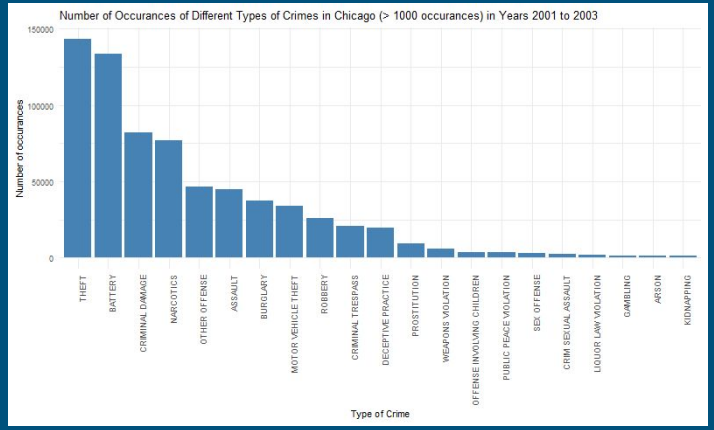
- Community District: The community district in which the crime occured
- Time: The time of day the crime occured
- Primary Type: The categorization of the crime
- Year: The year in which the crime was committed
- Latitude & Longitude: The latitude and longitude the crime was officially recorded at
- Community Area Name: Name of the community area in which the crime occured
- Per Capita Income: The per capita income for the area in which the crime occured
- Hardship Index: A score given to a community area based on the per capita income, population, area, and other factors

Thesis

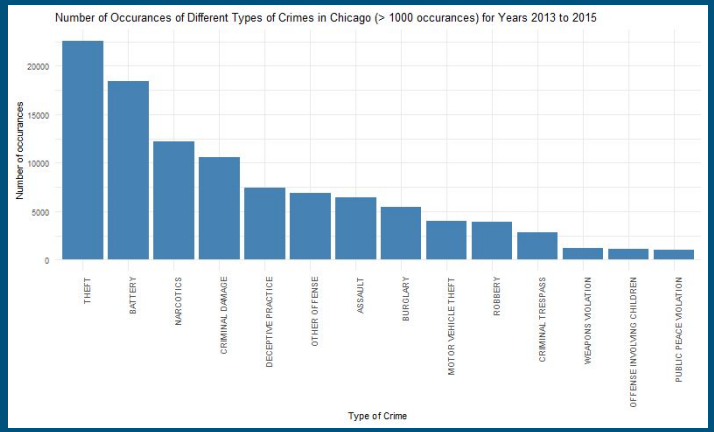
- Correlation between the household income, crime rate which is webbed with the type of community area people live in based on the income.
- Correlation between the hardship index, crime rate which is webbed with the type of community area people live in based on the income.

Data Cleaning

- We parsed the data into two groups
  - 2001 to 2003
  - 2013 to 2015
- We only chose to focus on the four top crimes committed in both sections:
  - Burglary
  - Motor Vehicle Theft
  - Narcotics
  - Robbery



2001 to 2003

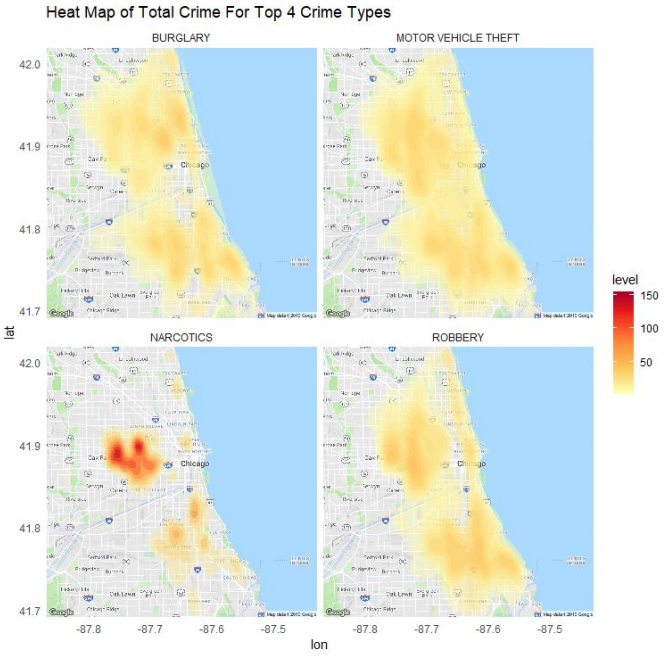


2013 to 2015

Exploratory Data Analysis (EDA)

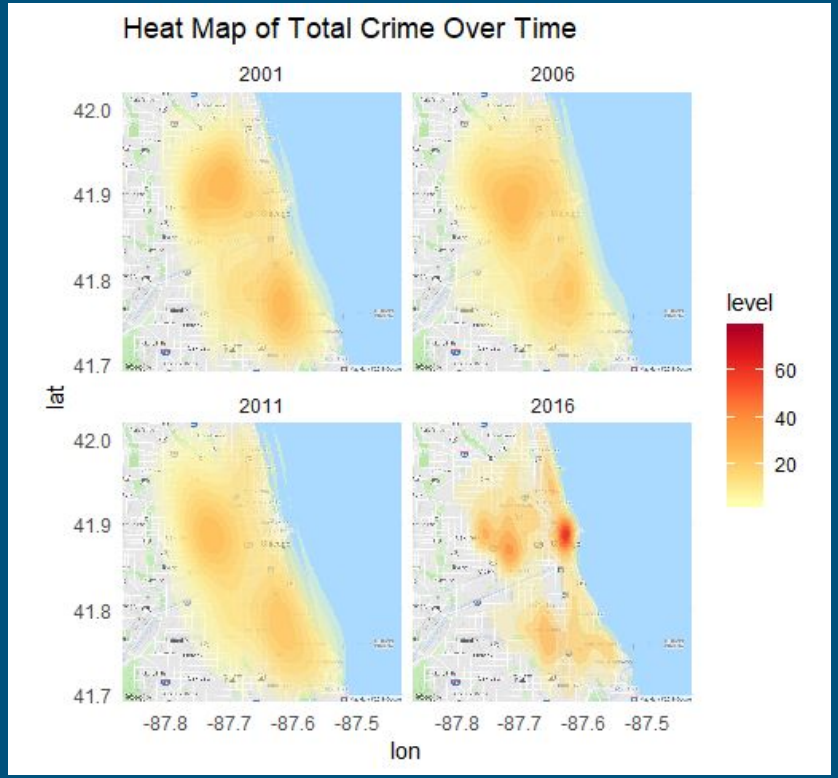
Plotting Data Using Google Maps API

- We first plotted a heatmap of the top four crimes in Chicago area
- We can see from the map that the spread of where crimes occur for Burglary, Motor Vehicle Theft, and Narcotics are very spread out and very similar to one another
- Narcotics crime is very concentrated in the inner city section of Chicago



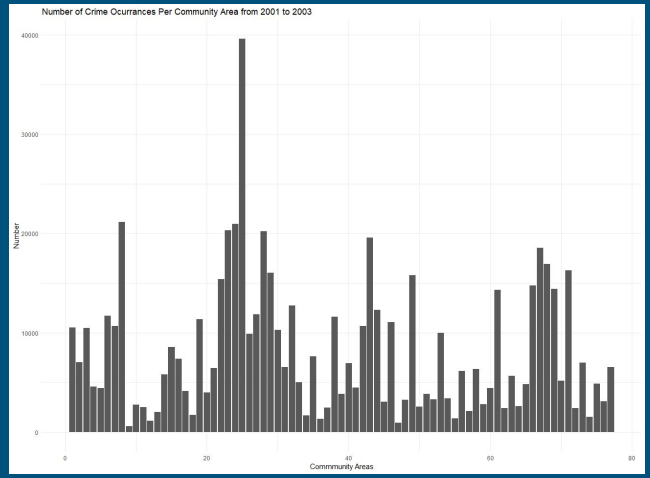
Plotting Data Using Google Maps API

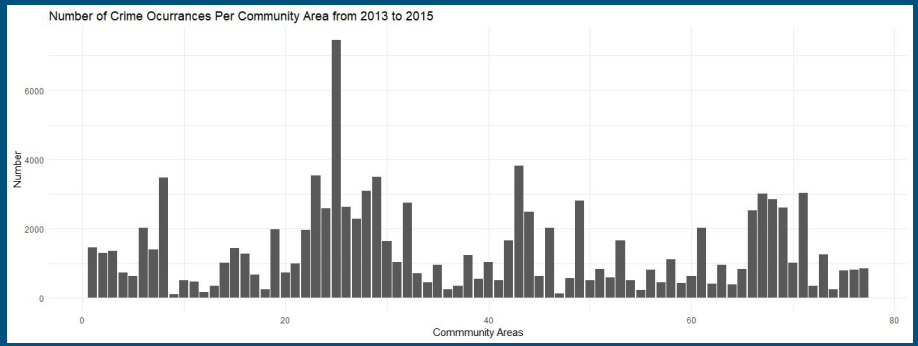
- We then plotted a heatmap of all crimes every 5 years, starting at 2001
- We can see that crime in 2016 is much more concentrated in certain areas



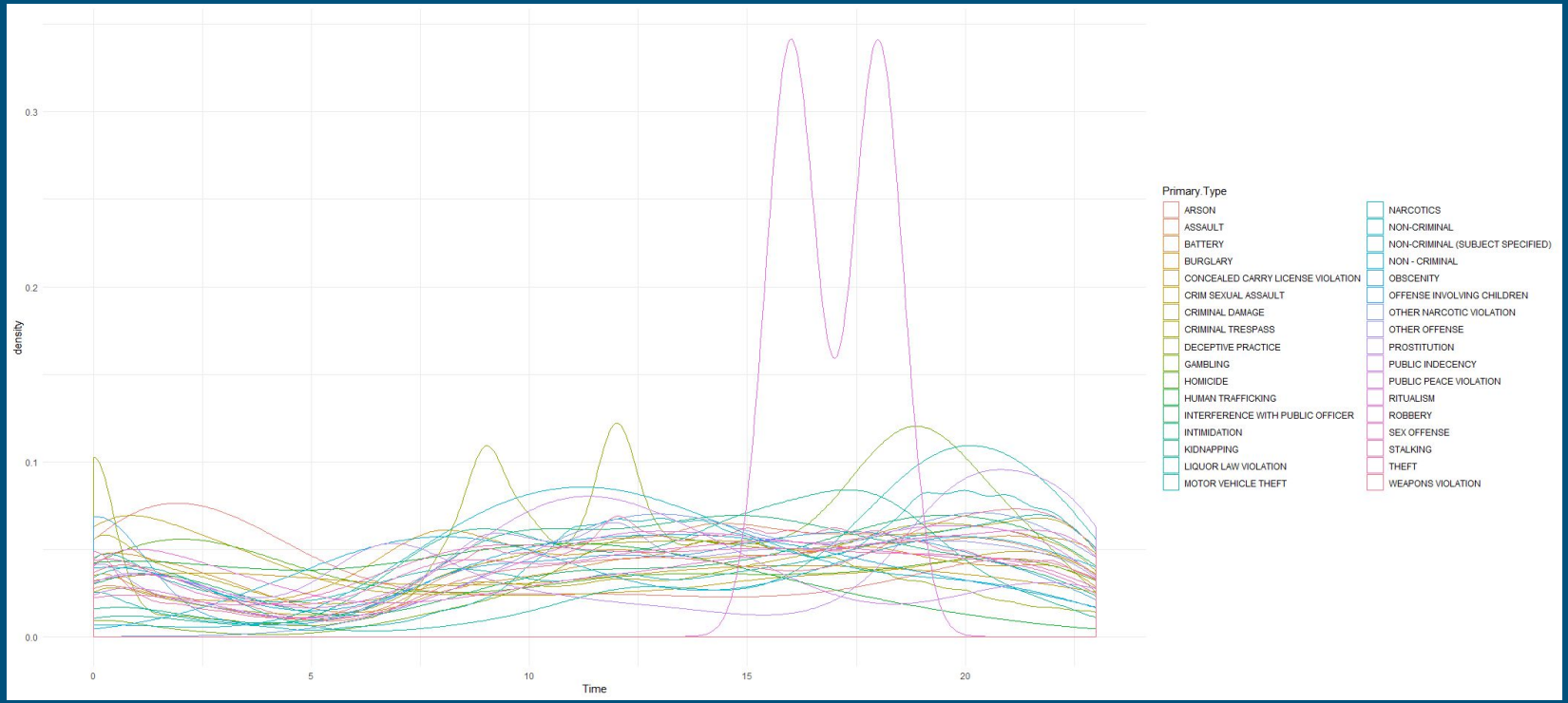
Distribution of Crime in Chicago Area

- The crime in Chicago is concentrated in one huge area
  - Austin
- It's an area that is so large that it emcompasses higher and lower income households
  - Some parts are safe
  - Other parts are concentrated with crime and violence

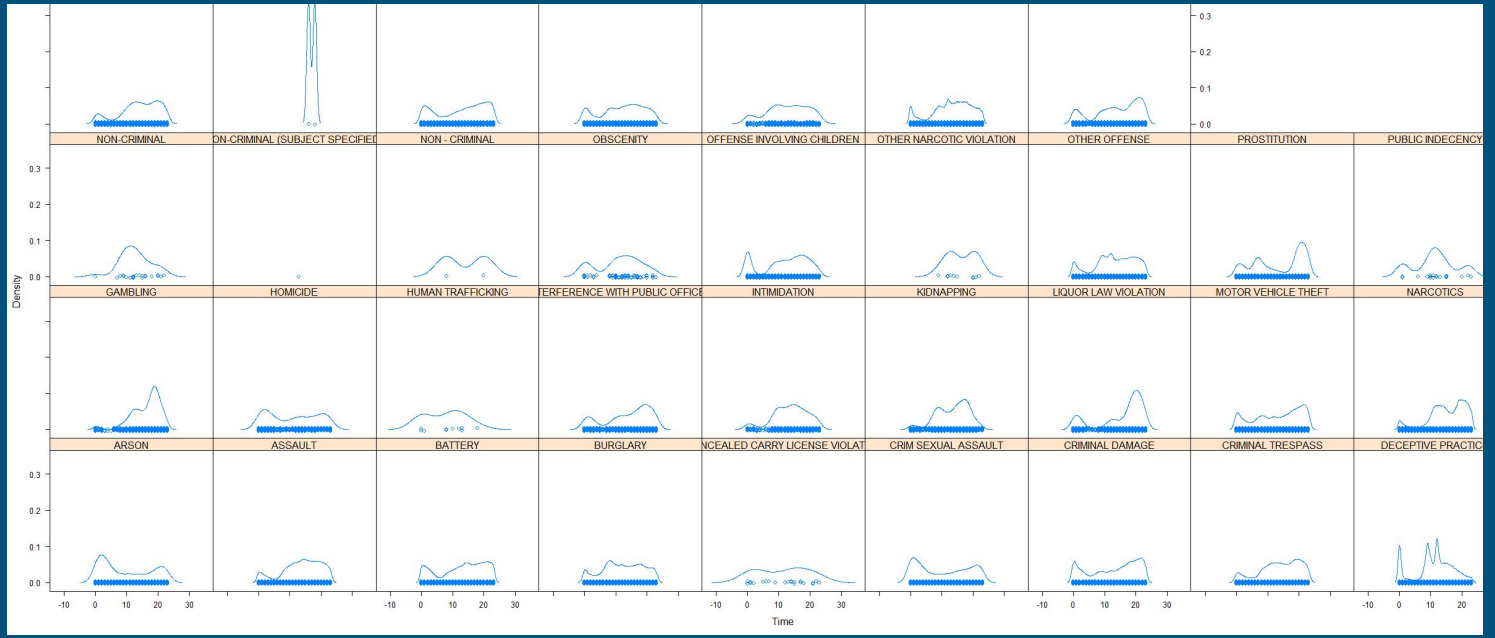




Density Plot of Time of Crimes



Density Plot of Time of Crimes



Results

Checking Assumptions

- Normal distribution: Fairly normally distributed, unimodal
- Independence: Each crime is independent of each other, no two can occur at the same time by same person
- Data is fairly representative of all crime in Chicago

2001-2003, Income

```
 Welch Two Sample t-test

data:  Mastervector$Number.1 and Mastervector$Income
t = -9, df = 100, p-value = 6e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21140 -13588
sample estimates:
mean of x mean of y
     8199     25563
```

- We performed a two sample t-test between Income and Crime Rate in the Chicago area between 2001 and 2003 and found that there is a significant difference. With a p-value of 6×10^{-15}, we reject the null hypothesis. There is sufficient evidence that there is relationship between income and crime rate in the Chicago area between 2001 and 2003.

2013-2015, Income

```
 Welch Two Sample t-test

data:  Mastervector$Number.2 and Mastervector$Income
t = -10, df = 80, p-value <2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -27655 -20692
sample estimates:
mean of x mean of y
     1390     25563
```

- We performed a two sample t-test between Income and Crime Rate in the Chicago area between 2013 and 2015 and found that there is a significant difference. With a p-value of 2×10^{-16}, we reject the null hypothesis. There is sufficient evidence that there is relationship between income and crime rate in the Chicago area between 2013 and 2015.

2001-2003, Hardship Index

```
 Welch Two Sample t-test

data:  Mastervector$Number.1 and Mastervector$`Hardship Index`
t = 10, df = 80, p-value <2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6622 9677
sample estimates:
mean of x mean of y
   8198.9      49.5
```

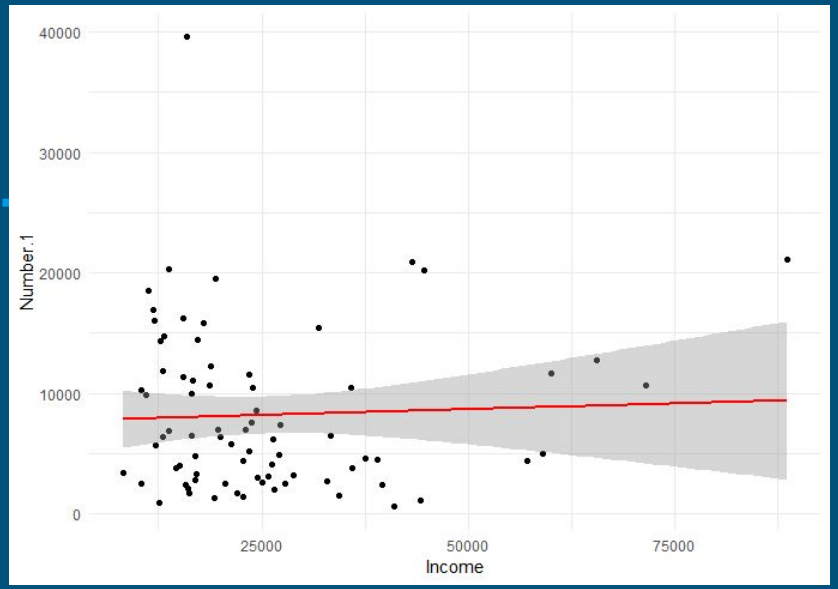
- We performed a two sample t-test between Income and Hardship Index in the Chicago area between 2001 and 2003 and found that there is a significant difference. With a p-value of 2×10^{-16}, we reject the null hypothesis. There is sufficient evidence that there is relationship between crime rate and hardship index in the Chicago area between 2001 and 2003.

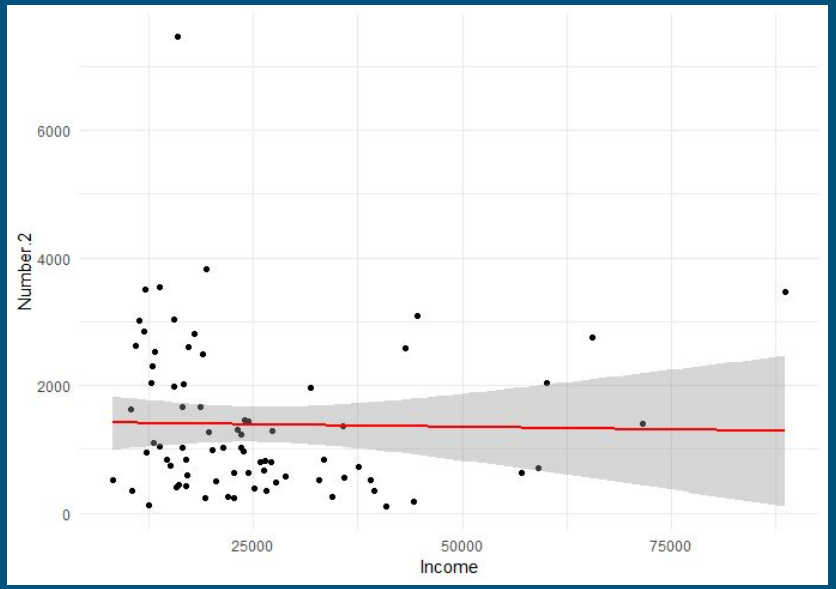
2013-2015, Hardship Index

```
 Welch Two Sample t-test

data:  Mastervector$Number.2 and Mastervector$`Hardship Index`
t = 10, df = 80, p-value = 5e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1066 1614
sample estimates:
mean of x mean of y
   1389.5      49.5
```

- We performed a two sample t-test between Income and Hardship Index in the Chicago area between 2013 and 2015 and found that there is a significant difference. With a p-value of 5×10^{-15}, we reject the null hypothesis. There is sufficient evidence that there is relationship between crime rate and hardship index in the Chicago area between 2013 and 2015.





- It’s intuitive to believe that as the per capita income increases, the crime rate decreases. However, these models contradict the belief and don’t depict a clear relationship. We believe that these linear regression models are not accurate because the points are very dense at first, then spaces out. There is not enough data for higher income households to accurately state a strong correlation between income and crime rates.

References

https://www.chicagotribune.com/news/local/breaking/ct-austin-population-decline-met-20170714-story.html

https://www.chicagomag.com/Chicago-Magazine/The-312/July-2012/Austin-Chicagos-Deadliest-Neighborhood/