



Crime Analysis - Chicago & Los Angeles

CIS5200-Group 2

Serena Tang, Alicia C Martinez,
Sophathriya Sen, Dmitrii Kim

Dec.4th
2024





Data and Server Specifications

 Project Idea Explanation

 Project Architecture and workflow

 Data Process

 Data Visualization and Analysis

Data and Server Specifications

Data Size and Source

Data Size: 2.9 GB

Data Coverage: 2020 -2024

Data Source URL:

<https://data.census.gov/table>

H/W experimental Specifications

Cluster version: AWS (EMR)

Cluster number of nodes: 5

Memory size: 30 GB

CPU speed: 2.5 GHz



Data and Server Specifications

Project Idea Explanation

Project Architecture and workflow

Data Process

Data Visualization and Analysis

Project Idea

- **Data Overview:**
 - Crime Data of Chicago city 2001-2024;
 - Crime Data of Los Angeles city 2020-2024;
 - Four Topic Census Data: education/ population / household income / unemployment
- **Project Originality:**
 - 1) The latest period after pandemic.
 - 2) Use spatial aggregate to validate crime data.
 - 3) Visualize crime data in ArcGIS online & map story
 - 4) Analysis the relationship between crime and four census topics.



Data and Server Specifications



Project Idea Explanation



Project Architecture and workflow

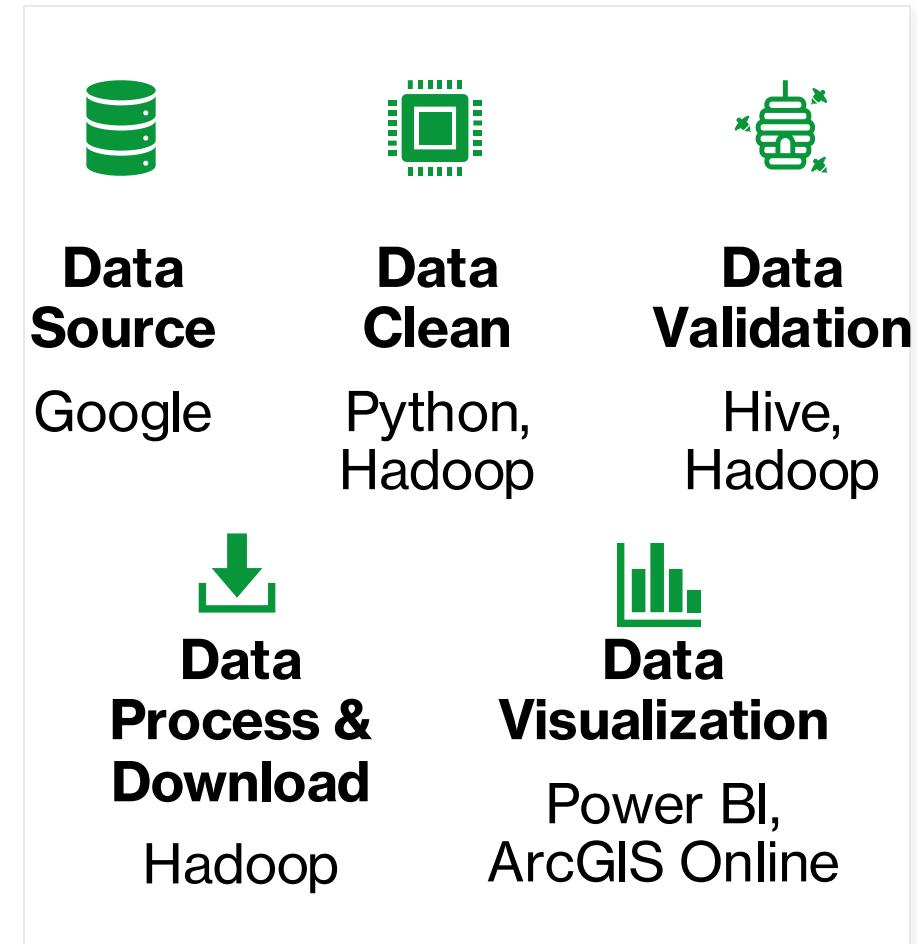
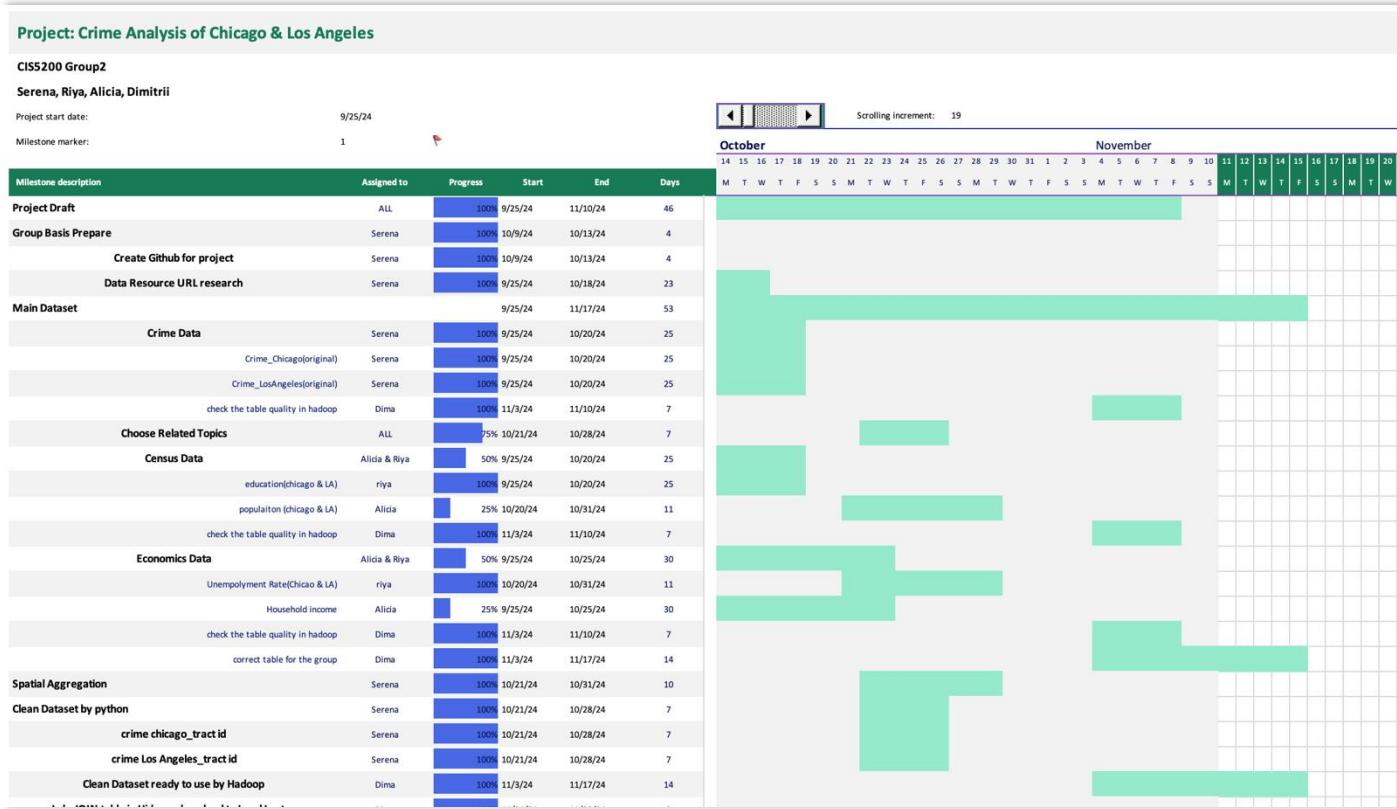


Data Process

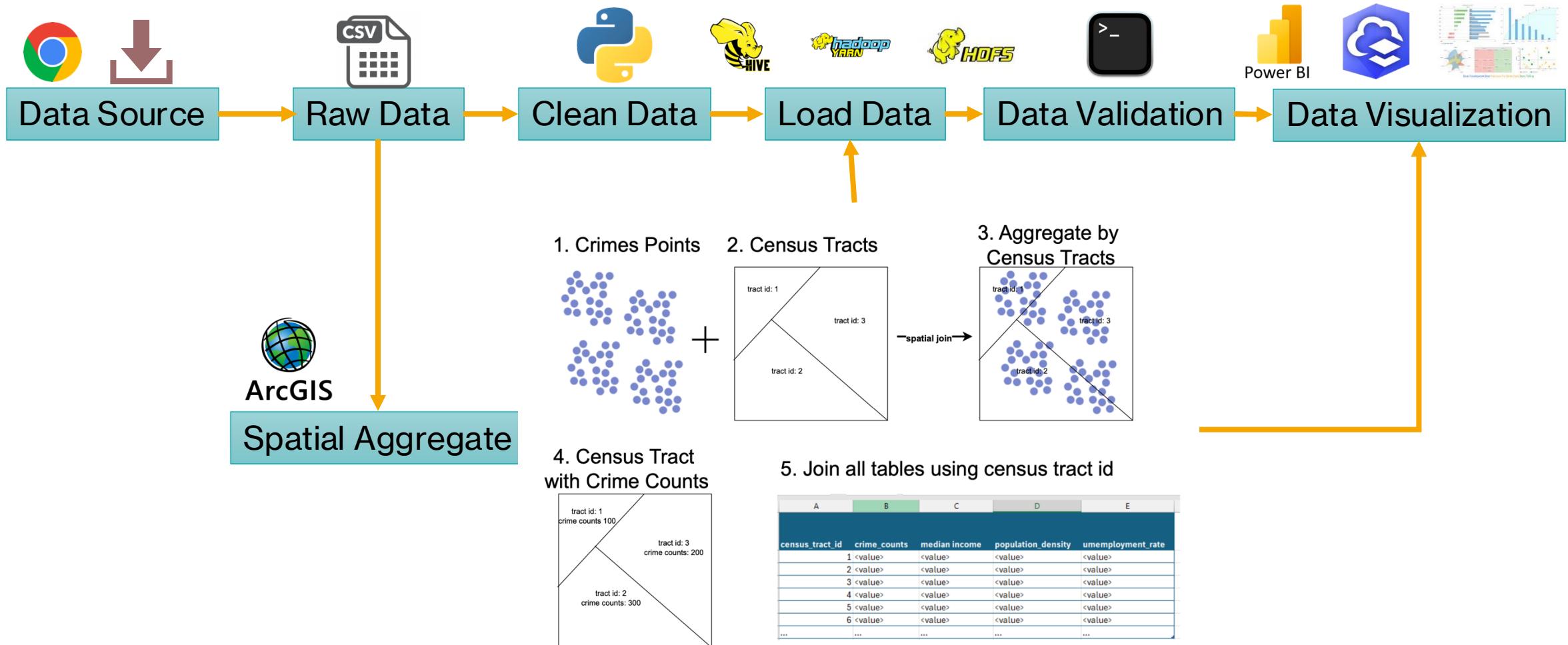


Data Visualization and Analysis

Project Architecture



Project Workflow





Data and Server Specifications

Project Idea Explanation

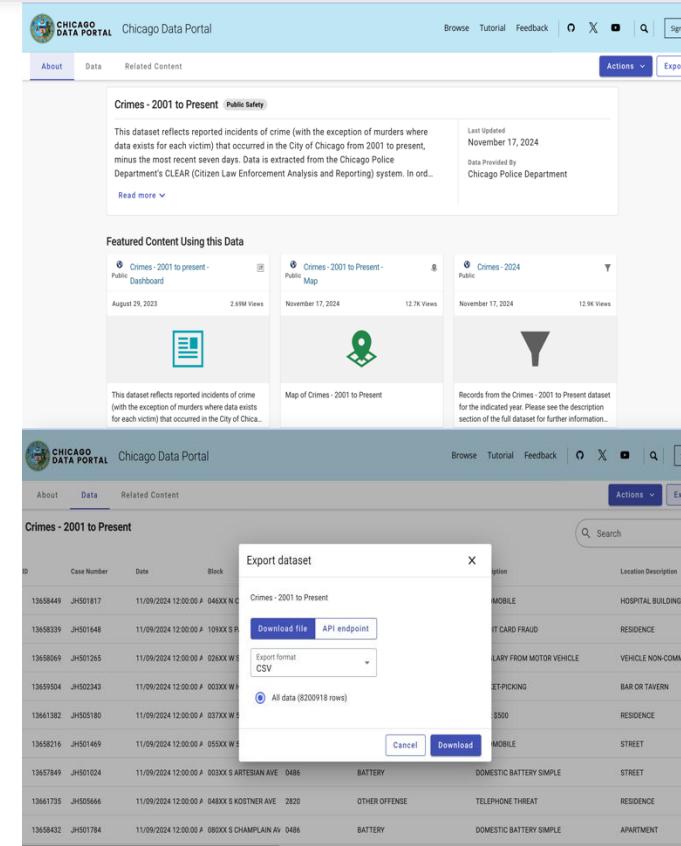
Project Architecture and workflow

Data Process

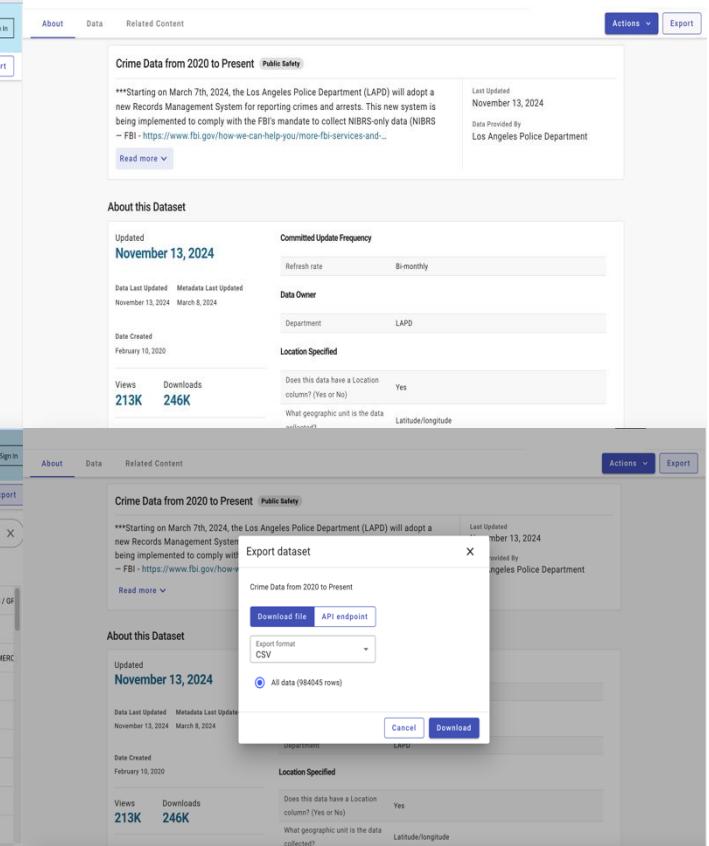
Data Visualization and Analysis

Data Process-Download Crime Data

- Steps
- **Crime Data**
- Go to the website for Chicago:
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data
- Go to the website for Los Angeles:
https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data
- Select Data: Export
 - export format: csv
- Download Data to Linux server and then to the Hadoop



The screenshot shows the Chicago Data Portal's dataset page for 'Crimes - 2001 to Present'. The page includes a brief description of the dataset, which reflects reported incidents of crime from 2001 to present, minus the most recent seven days. It is extracted from the Chicago Police Department's CLEAR system. Below the description are three featured content cards: 'Public Dashboard', 'Public Map', and 'Crimes - 2024'. A note below the cards states: 'This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In ord...'.



The screenshot shows the Los Angeles Data Portal's dataset page for 'Crimes Data from 2020 to Present'. It includes a note about the new Records Management System starting March 7, 2024. The page displays dataset details such as last updated date (November 13, 2024), data owner (Los Angeles Police Department), and location specified (Latitude/longitude). It also shows views (213K) and downloads (246K). At the bottom, there is a modal window titled 'Export dataset' with options for 'Download file' (CSV) and 'API endpoint', and a radio button for 'All data (820918 rows)'.

Data Process-Spatial Aggregate

Data

- **chicago_shapefile/**: Contains shapefiles for Chicago census tracts. ([download link](#))
- **la_shapefile/**: Contains shapefiles for Los Angeles census tracts. ([download link](#))

Usage

Install Dependencies:

```
pip install pandas/ geopandas / zipfile
```

Functions

aggregate_points

Aggregates crime points by census tract and type of crime.

```
def aggregate_points(points_gdf, geometry_gdf): points_gdf = gpd.sjoin(points_gdf, geometry_gdf, how='inner', predicate='within')  
points_gdf = points_gdf.groupby(['geoid10', 'Primary Type']).size().reset_index(name='count') return points_gdf
```

zip map data

combine and create shapefile by year and geoid and zip them.

Results

The results are saved as CSV file:

- crime_chicago_2001_2024_by_tract_type.csv
- crime_la_2020_2024_by_tract_type.csv
- crime_chicago_aggregated.zip
- crime_la_aggregated.zip

Data Process-Download Topic Data

- Steps
- Education
- Go to the website: <https://data.census.gov/all>
- Set filters:
 - Geographies → Census Tract → California → Los Angeles County → All Census Tracts
 - Topics → Education → Educational Attainment
 - Years → 2022
- Download zip
- Clean Data using Python
- Download Data to Linux server and then to the Hadoop
- **Repeat the same steps for Population, Income, Employment**

The screenshot shows the data.census.gov website interface. The top navigation bar includes links for 'All', 'Tables' (which is selected), 'Maps', 'Profiles', and 'Pages'. A search bar and an advanced search link are also present. On the left, there are three filter panels: 'Geographies' (set to 'All Census Tracts within Los Angeles County, California'), 'Topics' (set to 'Education'), and 'Years' (set to '2022'). The main search results are displayed under the heading 'Select Education'. A sidebar on the right lists '44 Results' from the American Community Survey, including items like 'B1501 | Educational Attainment' and 'B1502 | Field of Bachelor's Degree for First Major'. Each result entry includes a preview, a download link, and a 'Download Table Data' button.

Data Process-Clean Data

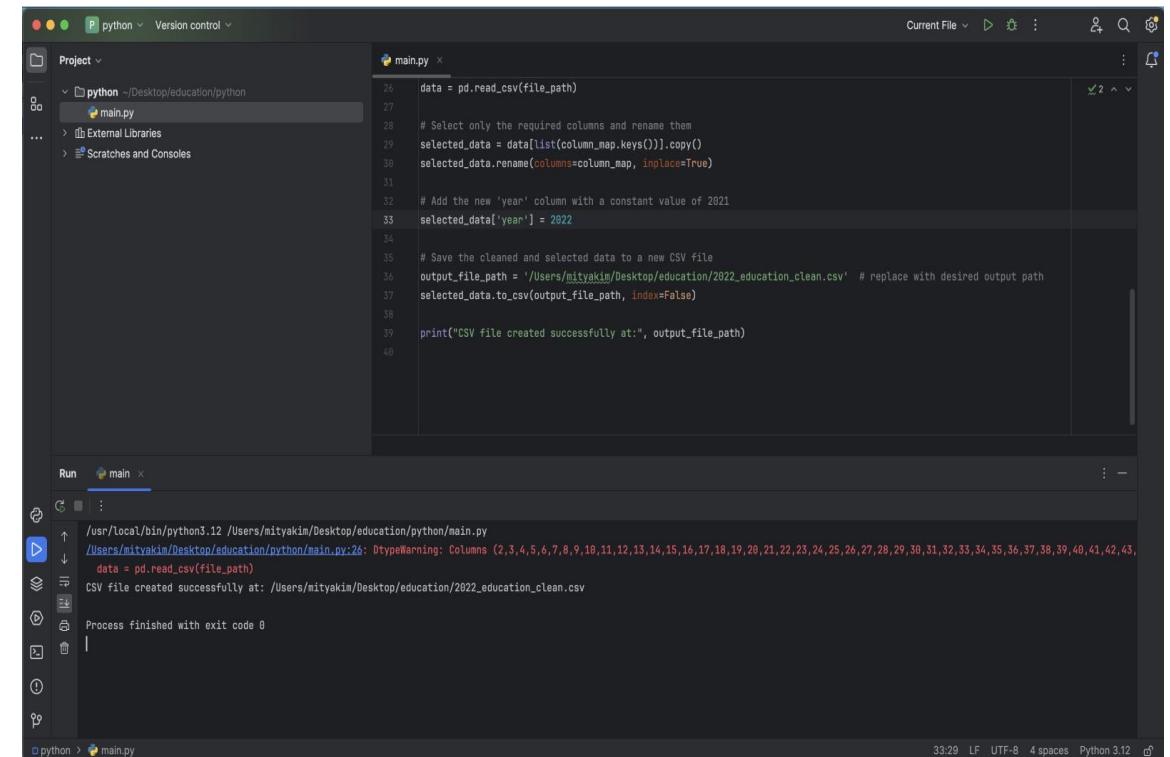
- Python language
- Import pandas library
- PyCharm IDE

```
column_map = {
    "GEO_ID": "geo_id",
    "NAME": "area_name",
    ...
}

# Select only the required columns and rename them

selected_data = data[list(column_map.keys())].copy()
selected_data.rename(columns=column_map, inplace=True)

....
```



```
data = pd.read_csv(file_path)

# Select only the required columns and rename them
selected_data = data[list(column_map.keys())].copy()
selected_data.rename(columns=column_map, inplace=True)

# Add the new 'year' column with a constant value of 2021
selected_data['year'] = 2022

# Save the cleaned and selected data to a new CSV file
output_file_path = '/Users/mityakim/Desktop/education/2022_education_clean.csv' # replace with desired output path
selected_data.to_csv(output_file_path, index=False)

print("CSV file created successfully at:", output_file_path)
```

Data Process-Hive & Hadoop

- Create original tables for each topic
- Join each topic with crime data using tractid and year as a key for a period from 2020 to 2022 years
- Clean tables
- Download final data to local PC and save it as .csv files for the future analysis

```
+-----+  
| tab_name |  
+-----+  
| chicago_education_and_crime |  
| chicago_employment_and_crime |  
| chicago_income_and_crime |  
| chicago_population_and_crime |  
| chicago_tractid |  
| crime_chicago |  
| crime_losangeles |  
| education_chicago |  
| education_la |  
| final_chicago_education_and_crime |  
| final_chicago_employment_and_crime |  
| final_chicago_income_and_crime |  
| final_chicago_population_and_crime |  
| final_la_education_and_crime |  
| final_la_employment_and_crime |  
| final_la_income_and_crime |  
| final_la_population_and_crime |  
| housing_data |  
| la_education_and_crime |  
| la_employment_and_crime |  
| la_income_and_crime |  
| la_population_and_crime |  
| la_tractid |  
| original_chicago_crime |  
| original_chicago_education |  
| original_chicago_employment |  
| original_chicago_income |  
| original_chicago_population |  
| original_la_crime |  
| original_la_education |  
| original_la_employment |  
| original_la_income |  
| original_la_population |  
| poverty_chicago |  
| poverty_la |  
| re_chicago_households |  
+-----+  
36 rows selected (0.253 seconds)  
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> █
```



Data and Server Specifications



Project Idea Explanation



Project Architecture and workflow

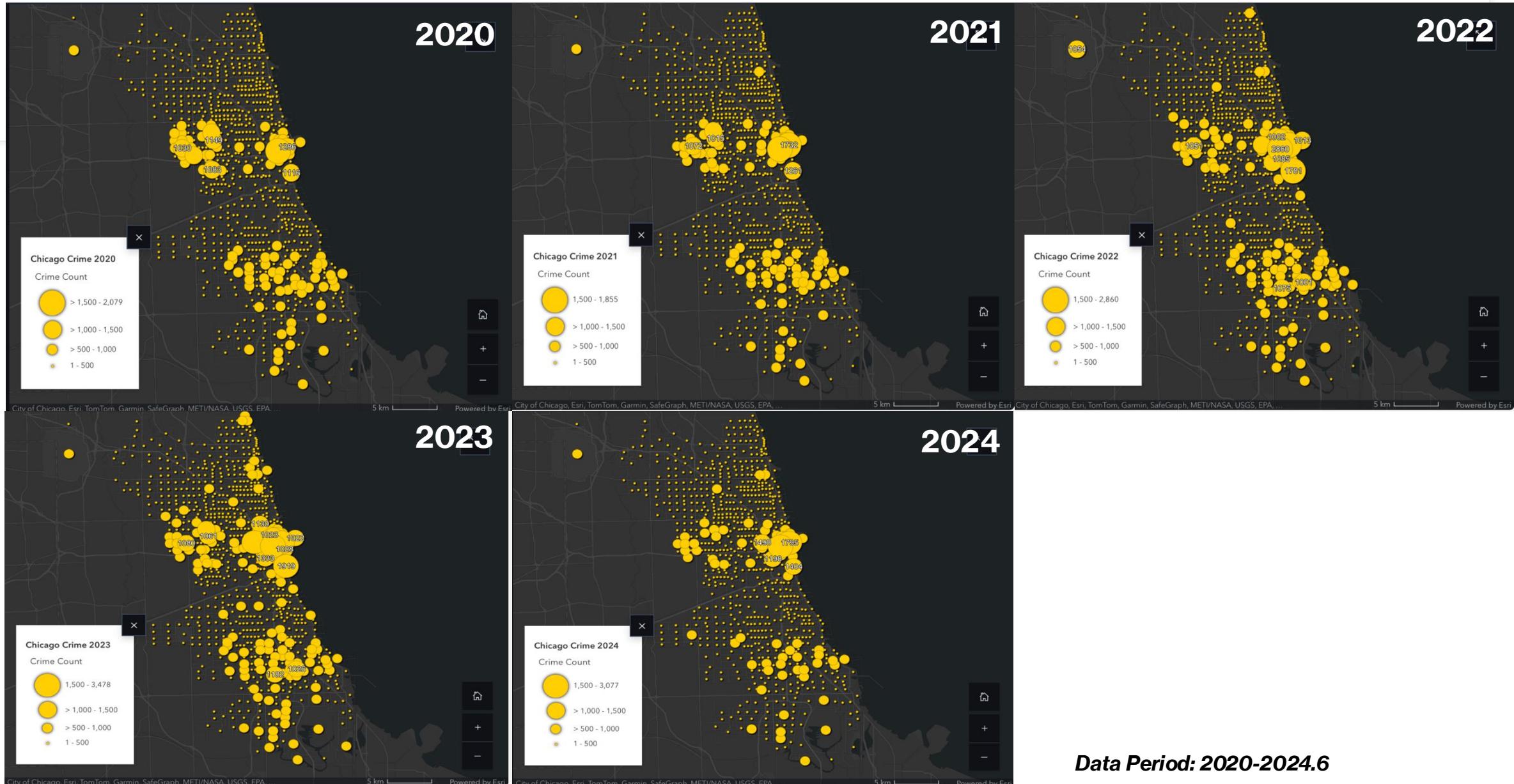


Data Process



Data Visualization and Analysis

Data Visualization & Analysis- Chicago Crime Map



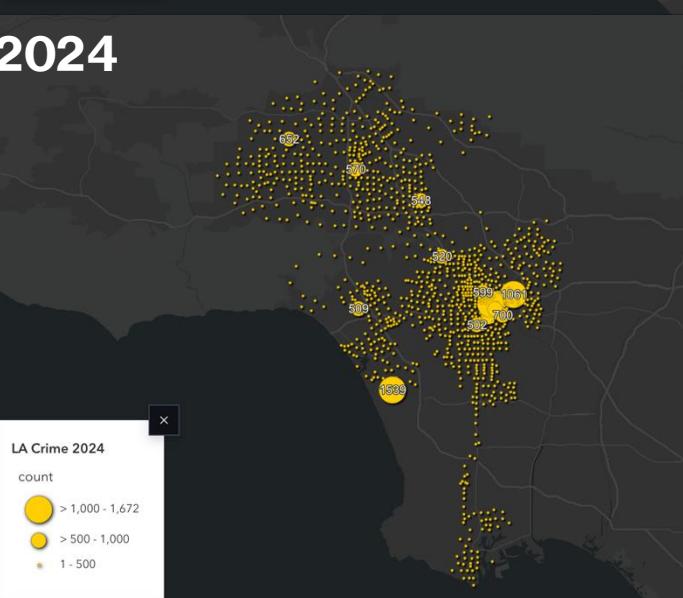
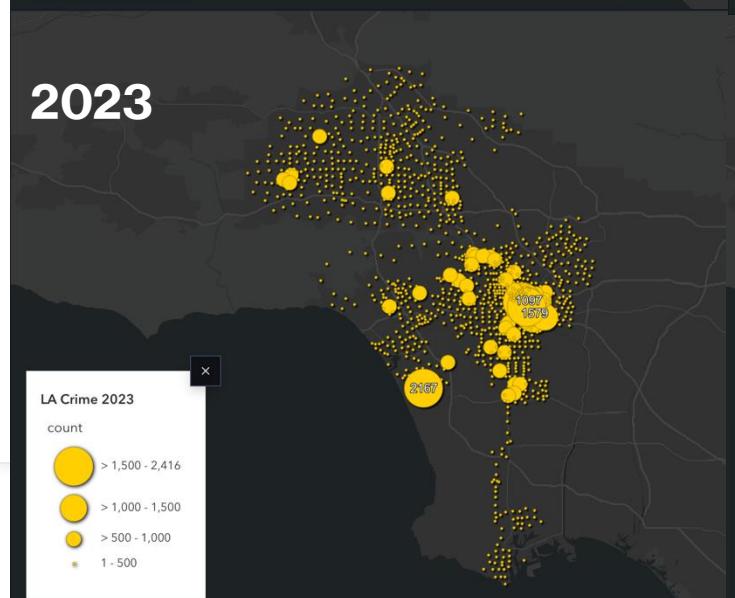
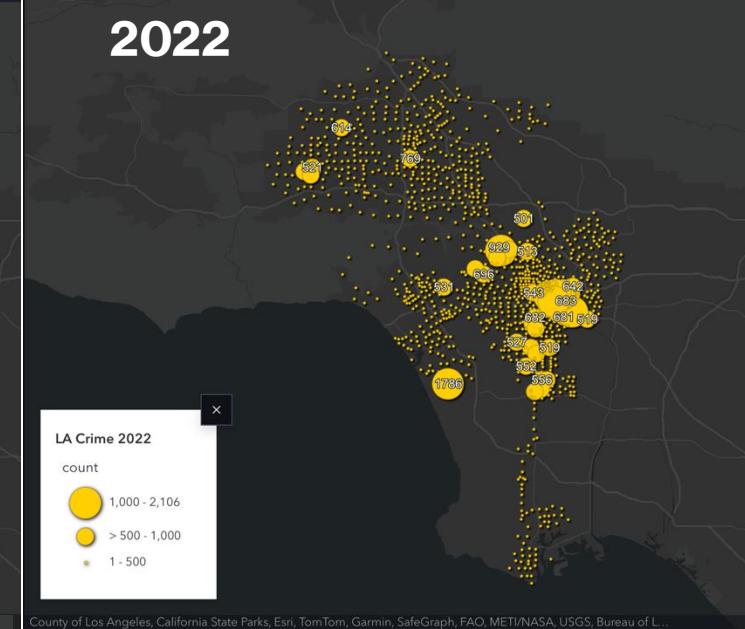
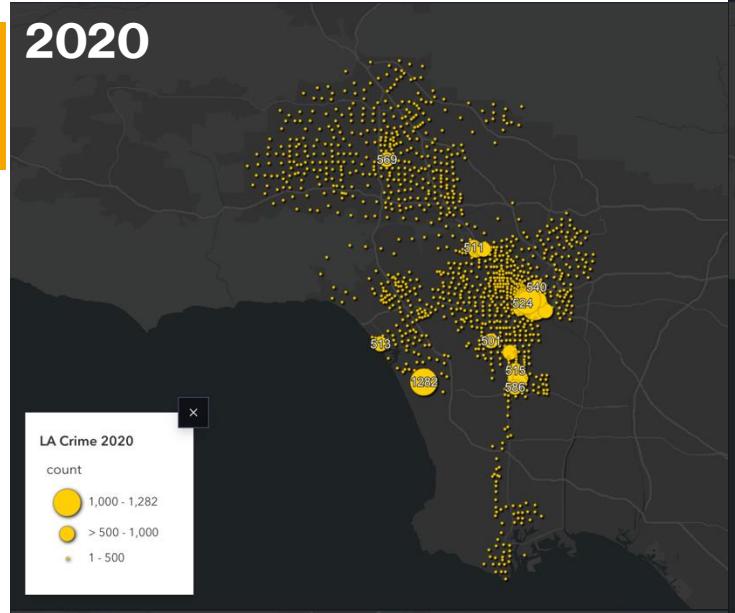
Data Visualization & Analysis

Chicago Crime Map



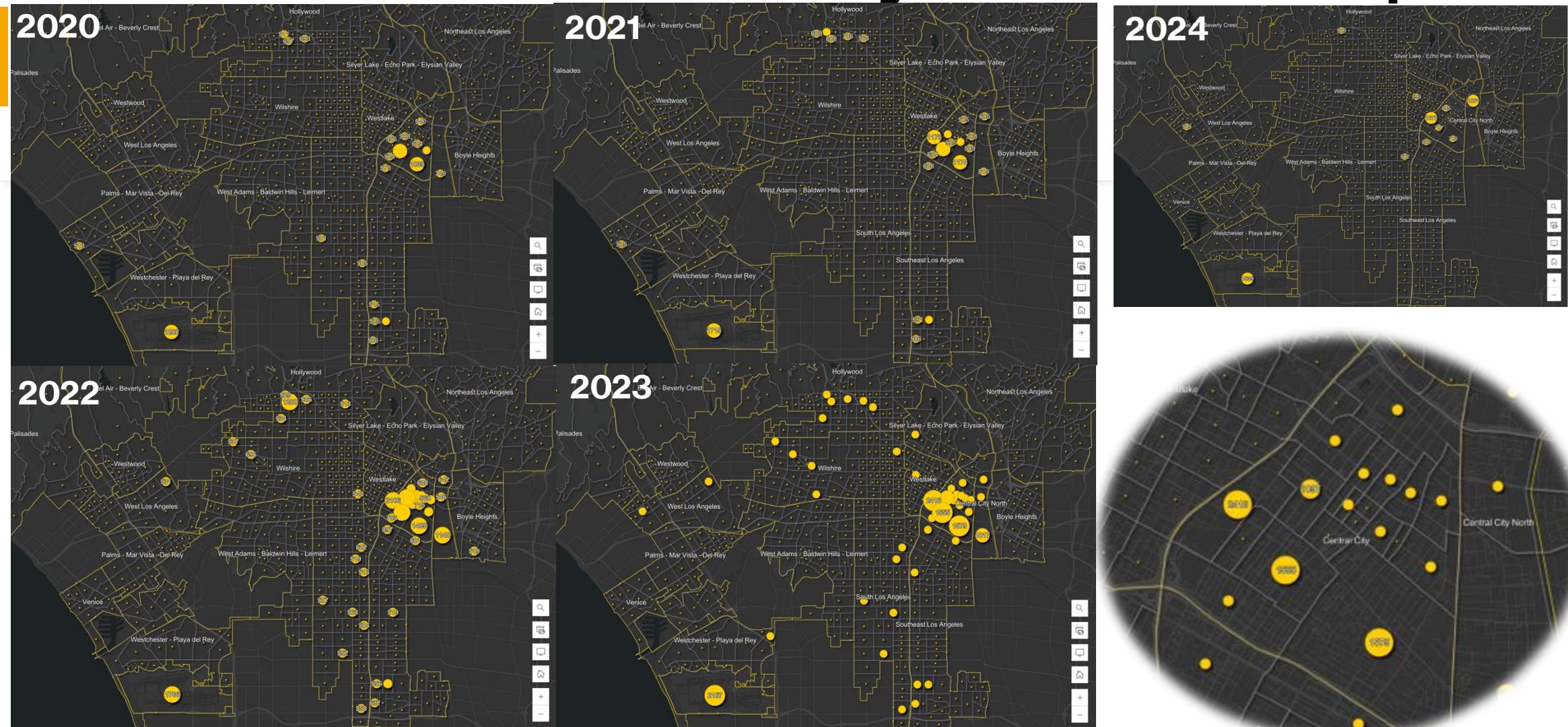
Data Period: 2020-2024.6

Data Visualization & Analysis-LA Crime Map



Data Period: 2020-2024

Data Visualization & Analysis-LA Crime Map

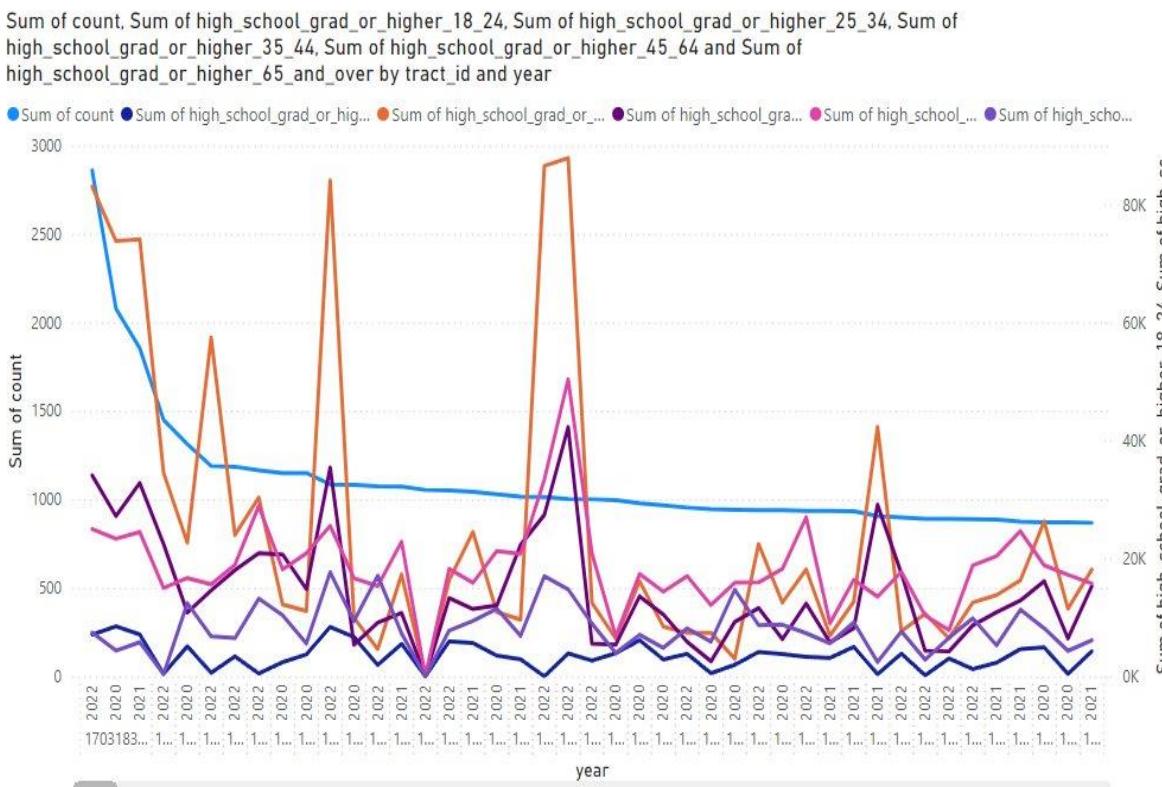


Data Period: 2020-2024

Data Visualization & Analysis

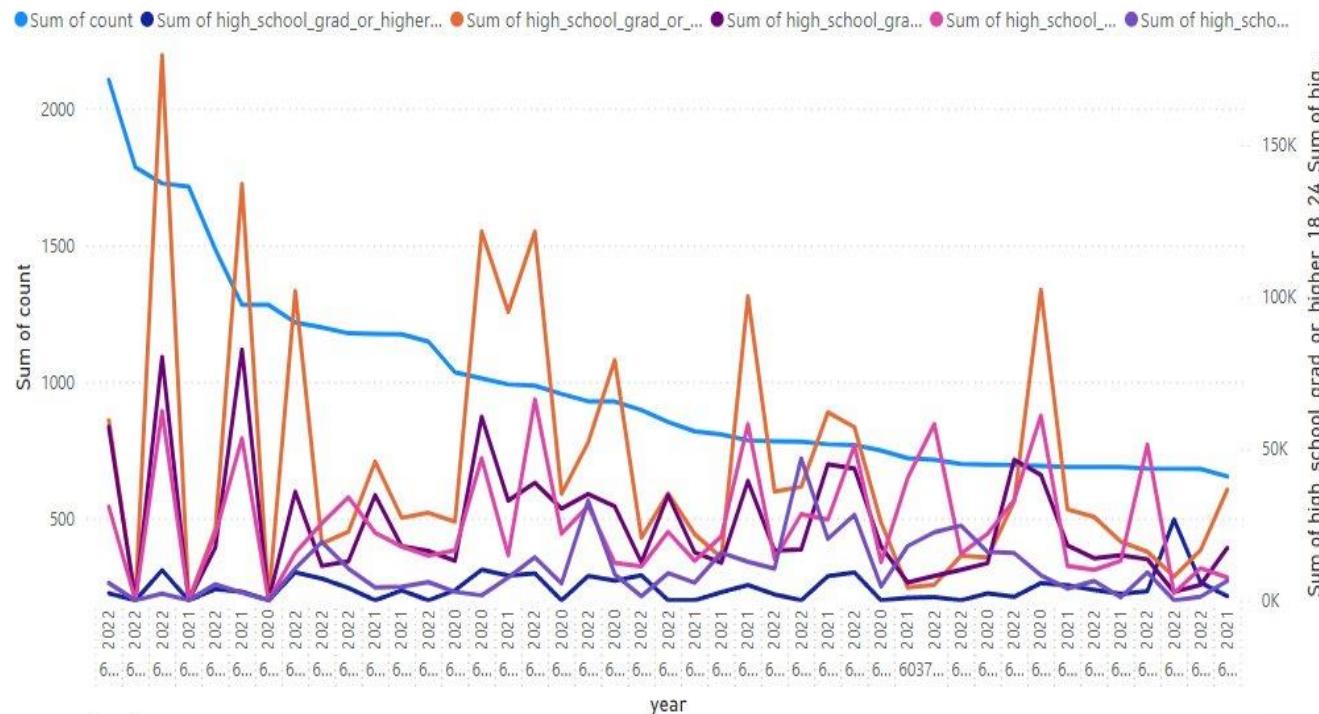
- Crime & Education Level

- Chicago



- LA

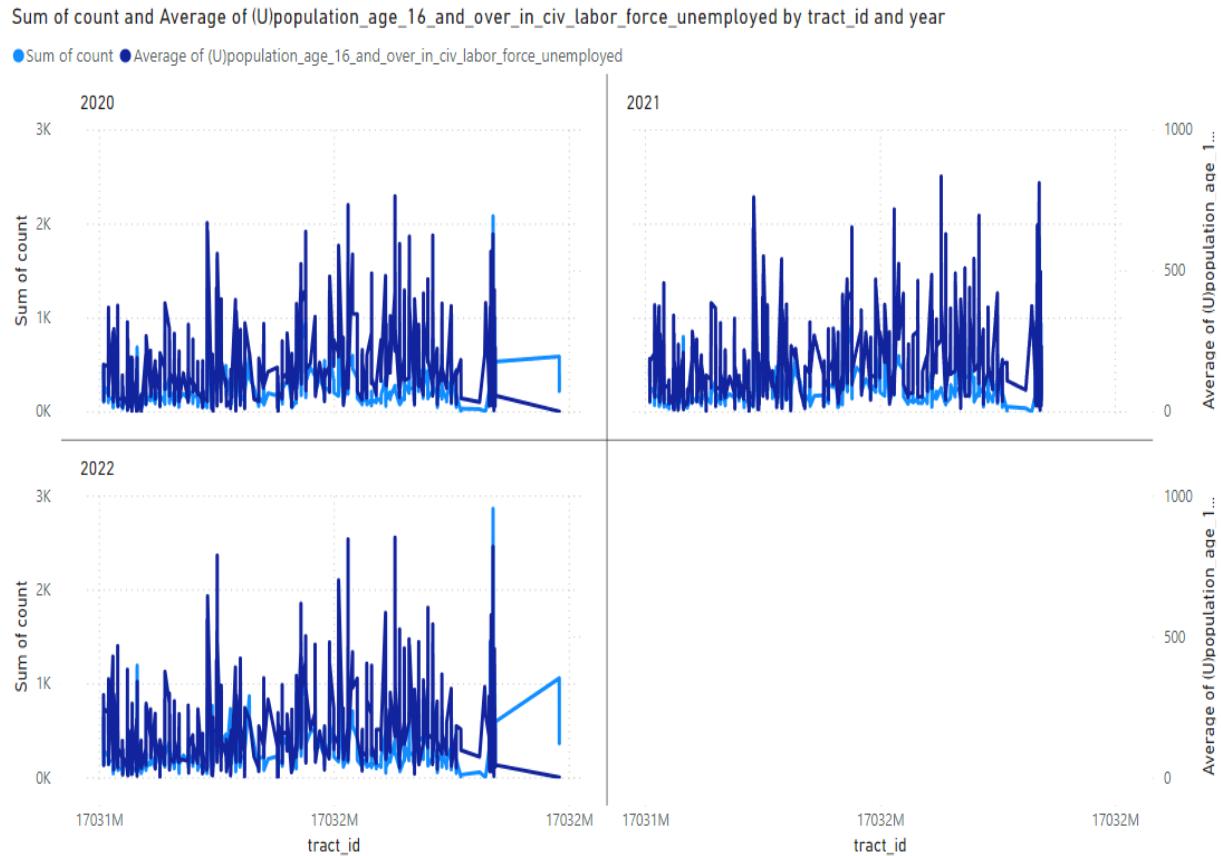
Sum of count, Sum of high_school_grad_or_higher_18_24, Sum of high_school_grad_or_higher_25_34, Sum of high_school_grad_or_higher_35_44, Sum of high_school_grad_or_higher_45_64 and Sum of high_school_grad_or_higher_65_and_over by tract_id and year



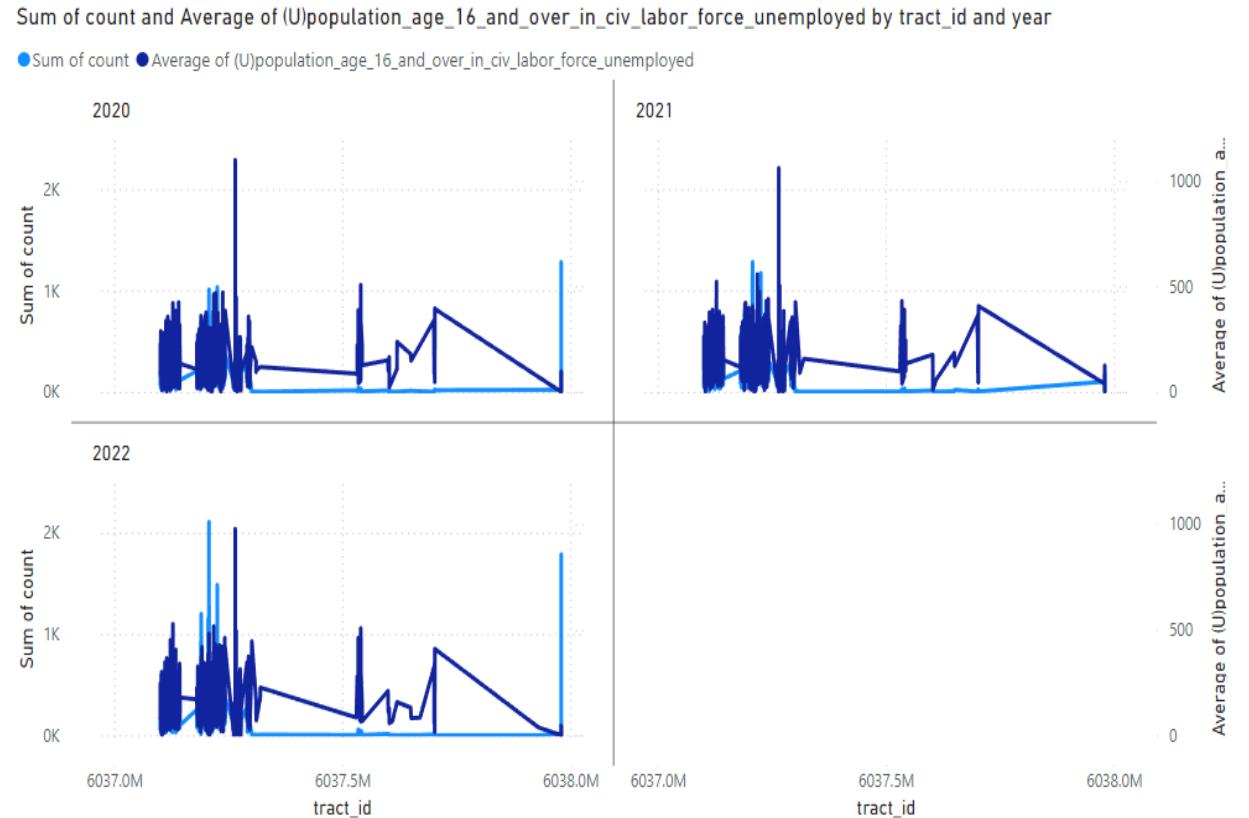
Data Visualization & Analysis

- Crime & Unemployment

- Chicago



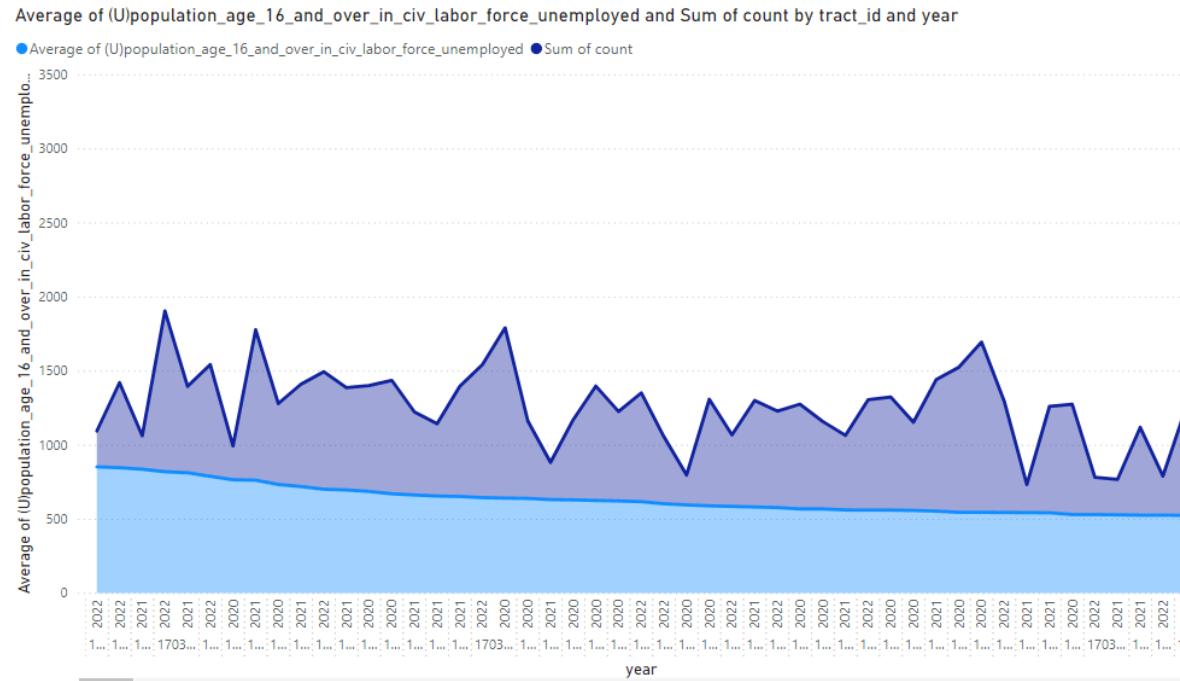
- LA



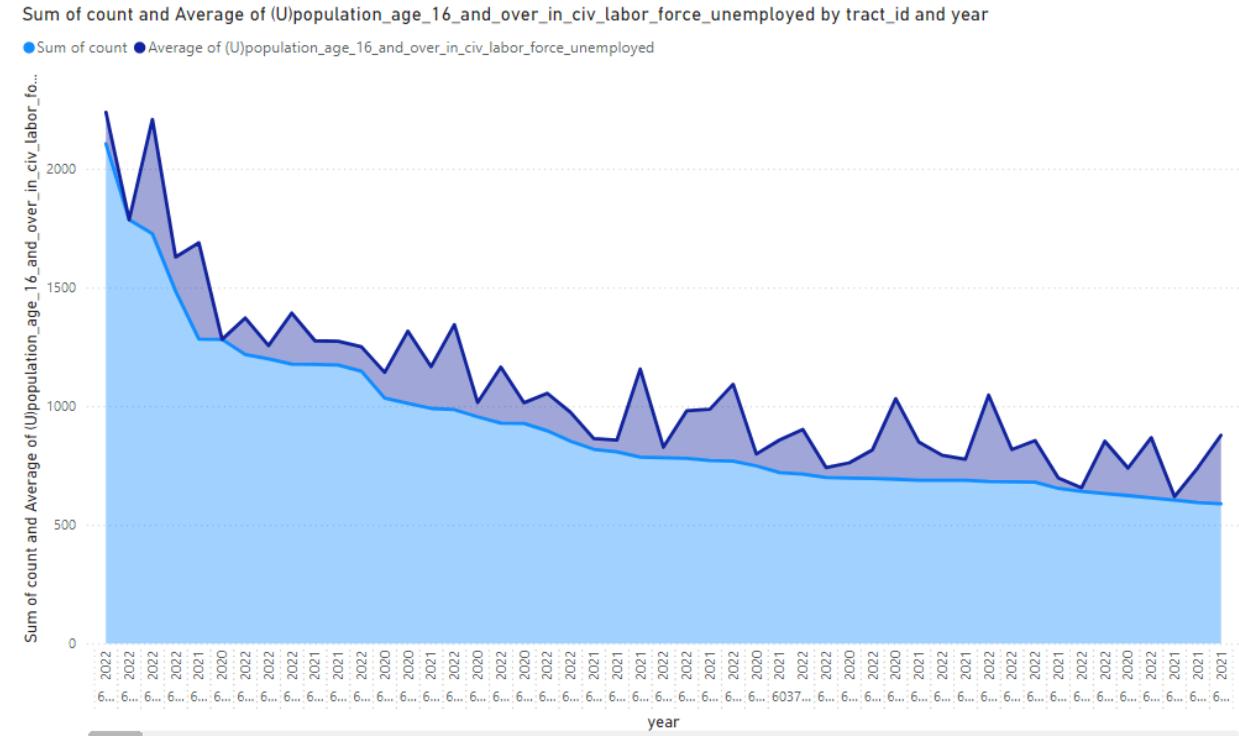
Data Visualization & Analysis

- Crime & Unemployment

- Chicago



- LA



Data Visualization & Analysis

- Chicago

- Crime & Population
- LA

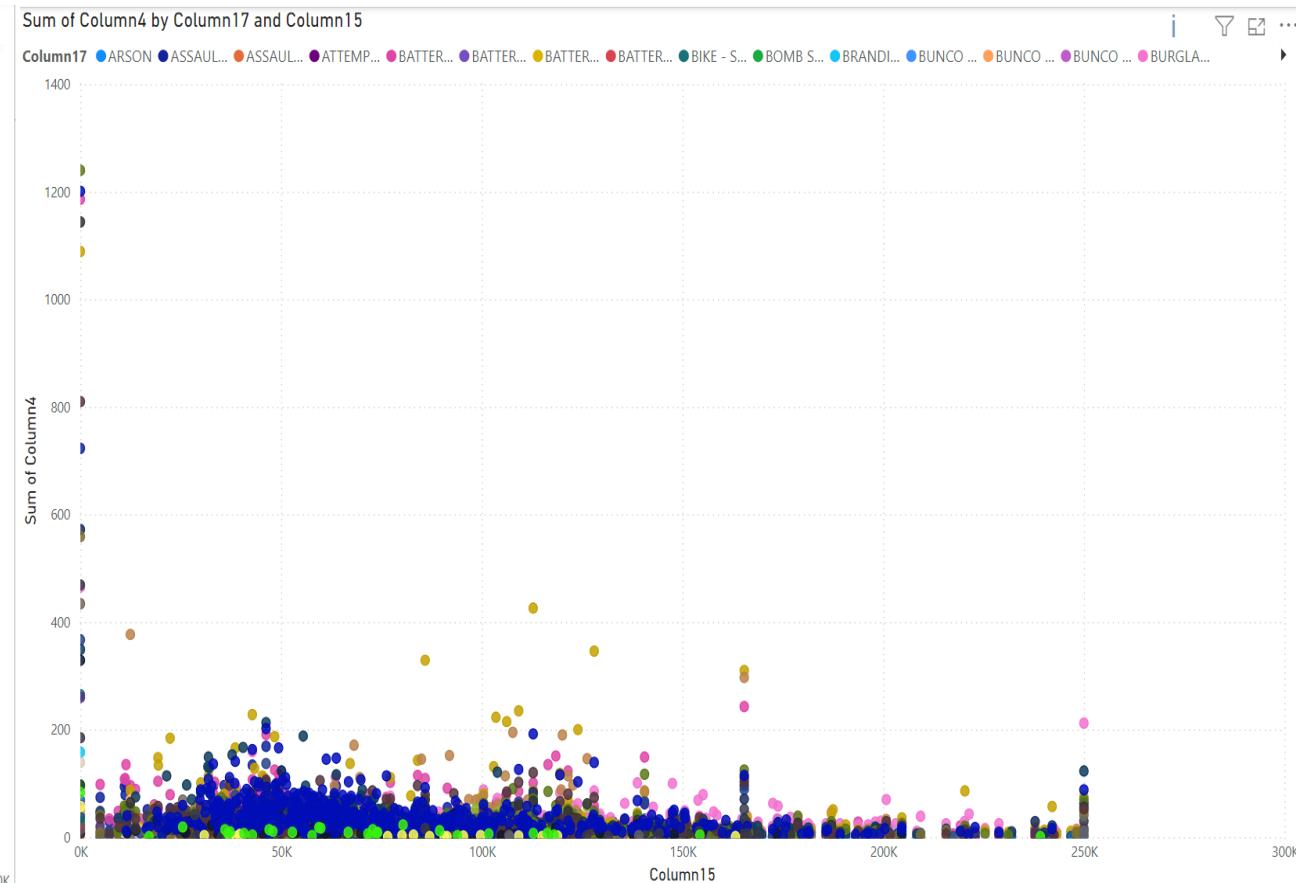
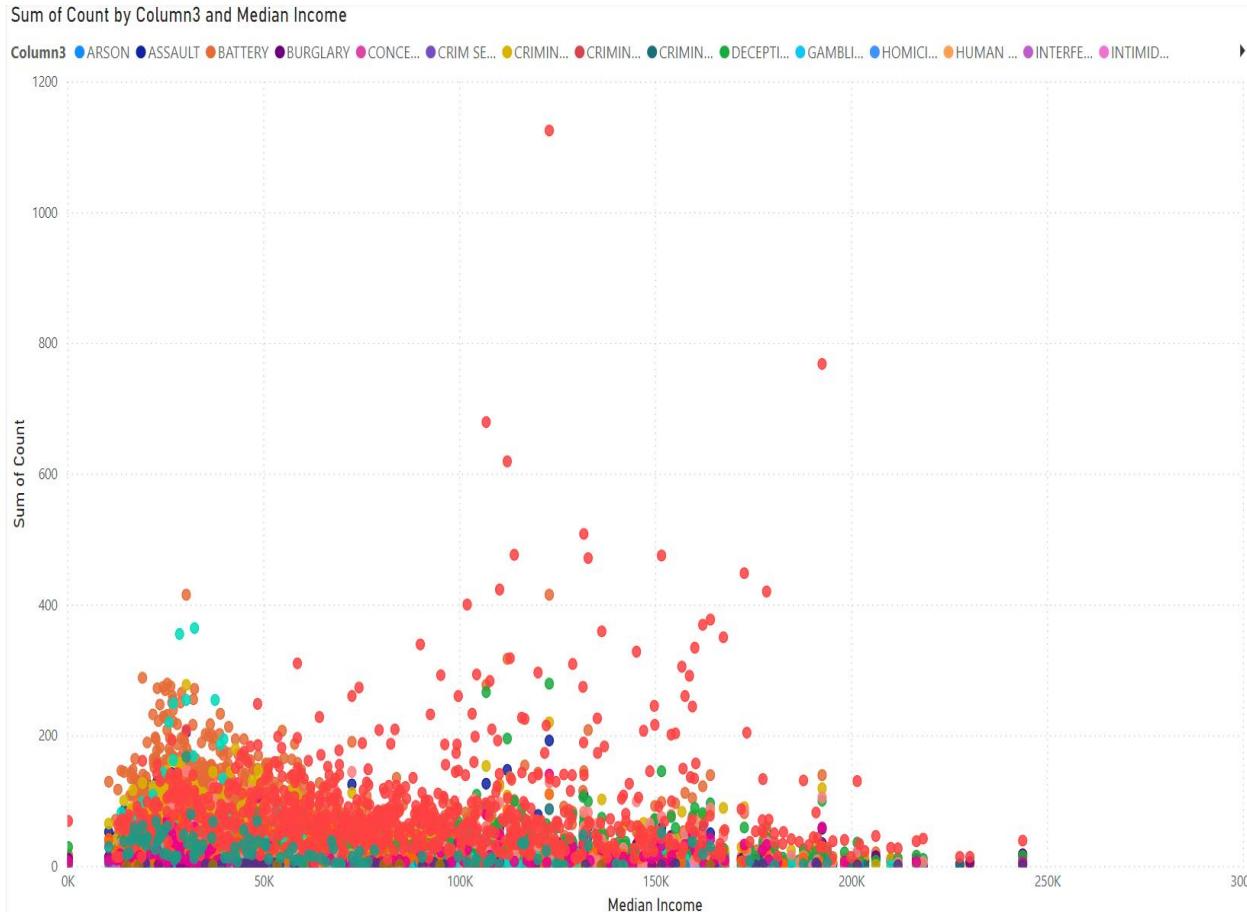
primary_type	Sum of african_american	Sum of American Indian/Alaska native	Sum of asian	Sum of chamorro	Sum of cuban	Sum of indian	Sum of japanese	Sum of native_hawaiian	Sum of pacific_islander	Sum of white	description	Sum of black or african american	Sum of asian	Sum of cherokee tribal grouping	Sum of chamorro	Sum of chinese	Sum of chippewa	Sum of eskimo	Sum of hawaiian	Sum of native_american	Sum of pacific_islander	Sum of white
ASSAULT	1538834		28279	132	49937	1891023	359982	115493	58942		ARSON	281707	283988		600	951	49645					
BATTERY	1537732		28392	132	50061	1900407	361950	116895	59131		ASSAULT WITH DEADLY WEAPON AGGRAVATED ASSAULT	680458	867343		1689	2565	151139					
BURGLARY	1531005		28157	132	49689	1875700	355752	114618	58269		ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER	125056	86081		106	293	19734					
CRIMINAL DAMAGE	1536136		28378	132	50010	1895516	361530	116895	58960		ATTEMPTED ROBBERY	431266	435576		818	1431	77369					
CRIMINAL TRESPASS	1496339		26962	132	47077	1783363	332396	102257	55826		BATTERY - SIMPLE ASSAULT	700525	934257		1742	2664	164298					
DECEPTIVE PRACTICE	1536102		28367	132	50037	1902801	362073	116908	59226		BATTERY ON A FIREFIGHTER	26928	34030		87	23	7938					
MOTOR VEHICLE THEFT	1535078		28293	132	49856	1874652	360214	116371	58825		BATTERY POLICE (SIMPLE)	238753	204791		277	1119	40236					
NARCOTICS	1411131		24108	132	38042	1501385	275591	88041	46807		BATTERY WITH SEXUAL CONTACT	365527	437296		836	1256	81021					
OTHER OFFENSE	1536553		28269	132	49898	1878853	360577	116754	58649		BEASTIILTY CRIME AGAINST NATURE SEXUAL ASSLT WITH ANIM	1878	385		0	0	184					
ROBBERY	1529128		27664	132	48544	1816232	348785	114887	55578		BIGAMY	276	160		0	0	3					
SEX OFFENSE	1124719		18237	132	33418	1218044	234367	65592	41044		BIKE - ATTEMPTED STOLEN	3481	3569		0	0	1381					
THEFT	1537799		28404	132	50088	1902959	362408	116908	59312		BIKE - STOLEN	300891	536331		916	1175	112158					
WEAPONS VIOLATION	1489485		26979	132	38520	1500499	294594	100808	47405		BLOCKING DOOR INDUCTION CENTER	630	430		9	0	31					
ARSON	853189		10891	106	13829	777035	106151	35317	15358		BOAT - STOLEN	10267	17986		23	13	2290					
OFFENSE INVOLVING CHILDREN	1396968		23566	103	33040	1300392	219059	62099	41850		BOMB SCARE	43466	68028		86	127	15709					
PUBLIC PEACE VIOLATION	1088817		13914	84	22753	1141505	175742	58908	25635		BRANDISH WEAPON	616639	724194		1385	2268	124432					
CRIMINAL SEXUAL ASSAULT	1259690		19961	76	36196	1244761	256913	82873	38158		BRIBERY	915	790		0	0	85					
HOMICIDE	1110301		11382	76	15816	812081	116958	45858	15535		BUNCO ATTEMPT	73145	47777		65	303	8659					
LIQUOR LAW VIOLATION	231993		5648	50	8058	320467	85870	31532	12476		BUNCO GRAND THEFT	427641	644046		1238	2049	114254					
STALKING	623509		9075	47	15186	455669	101159	29680	16046		BUNCO PETTY THEFT	252144	336545		698	1077	58854					
INTIMIDATION	277382		7207	35	11859	355187	88899	33619	11317		BURGLARY	690060	924309		1733	2572	163990					
INTERFERENCE WITH PUBLIC OFFICER	884795		8331	33	10043	751719	83497	29896	13030		BURGLARY ATTEMPTED	350017	475359		1180	1341	85387					
OBSCENITY	98882		2089	29	3017	128276	27993	10216	3189		BURGLARY FROM VEHICLE	690944	922079		1697	2680	163985					
KIDNAPPING	301901		3538	8	5761	252754	44479	17739	5619		BURGLARY FROM VEHICLE ATTEMPTED	109360	119114		266	606	22392					
CONCEALED CARRY LICENSE VIOLATION	275956		2754	0	2876	226768	33910	15805	3987		CHILD ABANDONMENT	7022	4788		0	0	1451					
CRIM SEXUAL ASSAULT	114939		616	0	4880	208885	21949	3164	2514		CHILD ABUSE (PHYSICAL) - AGGRAVATED ASSAULT	132106	76771		103	435	10612					
GAMBLING	53691		126	0	745	43756	9166	6062	1068		CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT	391879	339549		801	1448	53479					
HUMAN TRAFFICKING	11465		38	0	1022	10559	7150	1141	1423		CHILD ANNOYING (17YRS & UNDER)	138774	148011		267	651	25614					
NON-CRIMINAL	10787		52	0	82	1364	791	298	55		CHILD NEGLECT (SEE 300 W.I.C.)	173222	162414		417	427	26636					
OTHER NARCOTIC VIOLATION	11807		225	0	734	12756	7276	2544	959		CHILD PORNOGRAPHY	44954	44403		94	86	9930					
PROSTITUTION	121948		803	0	813	114398	12154	3872	1593		CHILD STEALING	116493	61369		265	244	10836					
PUBLIC INDECENCY	14827		228	0	403	23680	6498	2314	807		CONSPIRACY	2631	3727		0	19	271					
BITTIGATION	1042		0	0	0	3140	0	0	0		CONTUMACIE CONSPIRACY	3405	116167		561	477	16277					
Total	28084800	470933	2363	792290	31125955	5775833	1875364	928593			Total	25048390	30043448		57980	90914	5334900					

Data Visualization & Analysis

- Crime & Household Income

- Chicago

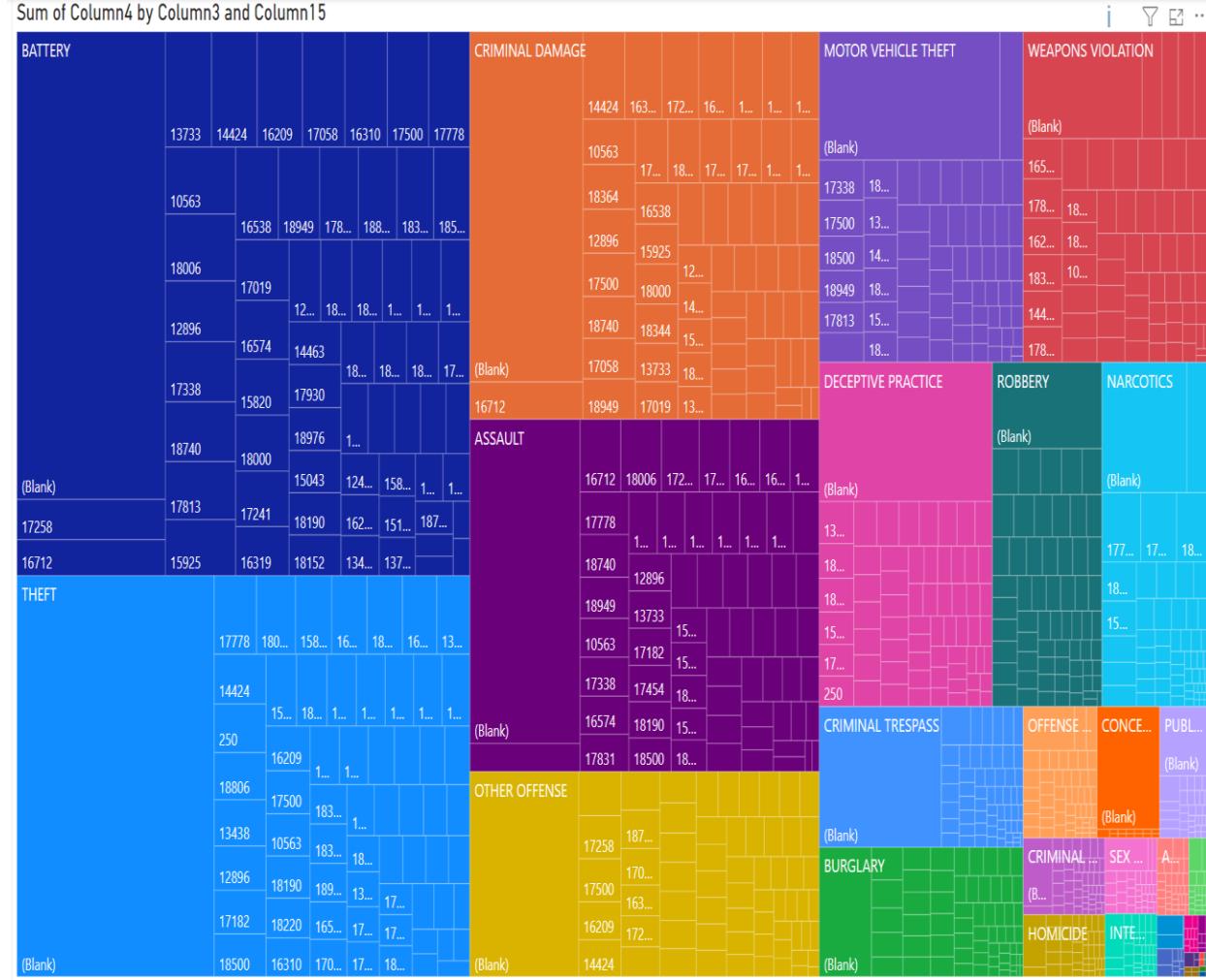
- LA



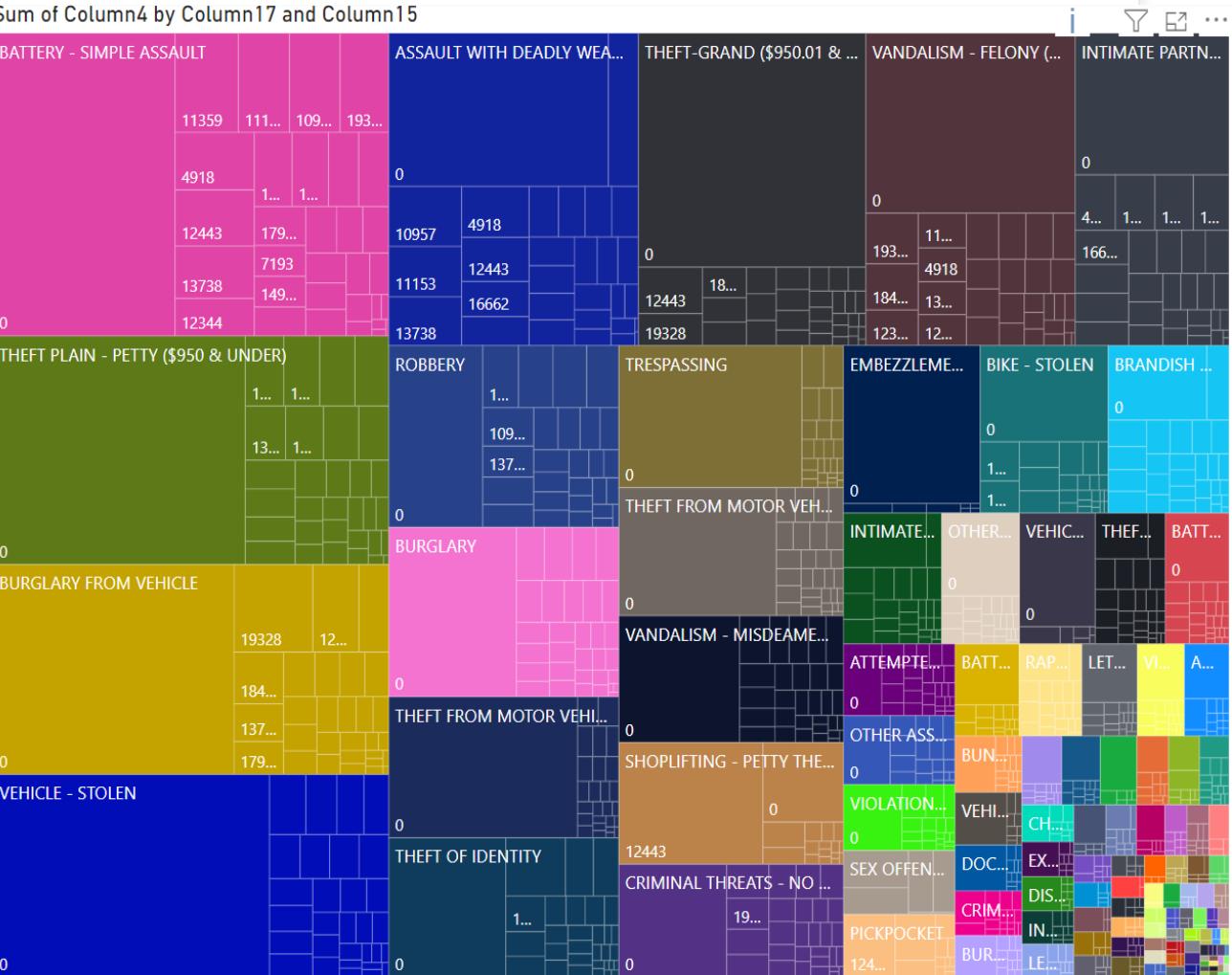
Data Visualization & Analysis

- Crime & Household Income

- Chicago



- LA



5200group2 GitHub
[https://github.com/
xuewentang/5200g
roup2/tree/main](https://github.com/xuewentang/5200group2/tree/main)

THANK YOU

chicago_shapefile	Chicago Shapesile	2 weeks
crime python	update LA crime clean python	6 hours
shapefile	LA shapefile	2 weeks
	update LA map data	6 hours
README.md	Update README.md	6 hours
csv	ADD new file	3 weeks
2001_2024_by_tract_type.csv	crime Chicago after clean	2 weeks
_2024_by_tract_type.csv	change type description to type code	2 weeks

pandas geopandas zipfile

Open and run the files in the `clean crime python/` directory (`crime_chicago.ipynb`, `crime_la.ipynb`) to analysis.

Points

points by census tract and type of crime.

```
def create_points(points_gdf, geometry_gdf):
    points_gdf = gpd.sjoin(points_gdf, geometry_gdf, how='inner', predicate='within')
    points_gdf = points_gdf.groupby(['geoid10', 'Primary Type']).size().reset_index(name='count')
    return points_gdf
```

Map data

Combine and create shapefile by year and geoid and zip them.

```
# Save the combined results to a shapefile
combined_results_2020_2024.to_file('map_data/crime_chicago_aggregated.shp')
```