DSCI 551: Foundations of Data Management　　　　　　　　**Due: April 25th, 2022**
Final Project Report: Data Science for Social Good
Yerkebulan Bauyrzhanov, Zihao Han, Soumeya Kerrar

Project Title: Social Inequality Across Los Angeles County

Code Files: https://drive.google.com/drive/u/1/folders/1QQflG4mvLVX3ZE1qiTvRwQffg4IK_50Y

GitHub Repository: https://github.com/bauyrzha/DSCI551-project

Application Link: https://share.streamlit.io/bauyrzha/dsci551-project/main/main.py

***Topic*:**
This project examines inequalities in regions of Los Angeles based on crime rates, access to food, and access to quality education**.**

***Motivation*:**

With the abundance of data available in today's age, data science is a tool that can be yielded to make significant contributions to society. One of today's most pressing issues, both domestically and abroad, is social inequity. Specifically, income inequality and public safety threaten the quality of life of many individuals. We can see this right here in Los Angeles County, with issues such as housing affordability, homelessness, food insecurity, crime rates, education access and others permeating the county. In this project, we aim to address some of these issues in order to provide users with awareness about specific inequalities and issues, and offer information and solutions to these issues. The purpose of this project is to build an app that compares different inequalities in different areas of the county in order to determine which areas need most improvement, funding, resources and public safety. With income inequality continuing to grow in this area of the state, it is important to examine who is most impacted in order to know what types of changes need to be made. This app will help people by providing necessary information about criminality, food availability, and education accessibility in LA county. By understanding existing inequality in LA, we can better help our city and neighborhoods especially at this special time. The app will also provide information on resources and initiatives currently available to address these issues. Users will be able to search by zipcode to see which areas are most negatively impacted, and also filter by categories of "crime", "education" and "food".

Application:

-　Architecture**:**
Our application was built on Streamlit. A Github repository is used to store the main script, the dataset of crime and other files (requirements.txt, ReadMe). Firestore and Firebase are used to store food and education datasets respectively. All code was written in Python.

-　Components**:**
The application is divided into three main components: crime in Los Angeles, food banks in Los Angeles and education in Los Angeles. Each of these components includes a geospatial visualization of where crimes occur, food is located, and schools are located. For the crime component, additional analysis on the number of crimes per hour, number of crimes per zipcode, and other crime-related statistics are offered. For the food component, a map of food banks and Grocery outlets is generated. A link to CalFresh application and information, which

offers financial assistance for purchasing food is also provided. For the education component, it provides a visualization of the density of schools. The school query is based on the relevant tag or zipcode area, and there is also a visualization of the density of the query results. There is also an integrated widget at the end for performing combinatorial analysis visualizations for discovering and drawing conclusions.

- Data Flow:

The data we collected gives us an idea of what crimes are occuring in the city, what areas are considered food deserts or have low access to food, and where high quality or poor quality schools are located. The crime dataset is downloaded directly from the lacity.org database. Both food and education data are scraped using web pages, which are cleaned and uploaded to a cloud database. The food datasets are uploaded to Firestore, and the education dataset is uploaded to Firebase. For each component and page for crime, food and education, we have defined functions which read the data from these different sources to display the analysis on the pages.

- **Datasets:**

Crime:
LA City Crime Data: https://data.lacity.org/browse?q=crime%20data&sortBy=relevance

Description: This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy. This data is as accurate as the data in the database.

Features included:
- Number of crimes by hour
- Number of crimes in context of zipcodes
- Locations of highest crime rates
- Locations of lowest crime rates
- Types of crimes(ranked more common and less common)
- Victim Sex

Variables:
dr_no - Division of Records Number: Official file number.
date/time - date and time crimes occur.
crm cd desc - definition of crime
vict age - victim age
vict sex - victim sex (F - Female M - Male X - Unknown)
location - address where crimes occur
lat - latitude
lon - longitude
zipcode - zipcode

Data extracting and data cleaning:
1. download dataset and clean/remove unnecessary/irrelevant data.
   run jupyter notebook **crime_getting_and_cleaning.ipynb**
2. get zipcodes for locations of crimes using uszipcode library in python. Need to use google colab.

run jupyter notebook **crime_get_zip_google_colab.ipynb** on google colab.

Data storage:
The extracted and cleaned crime dataset is stored in GitHub Repository due to it containing more than 465K rows which makes it impossible to keep it in Firestore or Firebase cloud databases in free plans.

Food Access:
We used the foodpantries.org list of food banks as well as the Grocery Outlet Store Locator, as this is considered an affordable grocery option, to give a general idea of food resources across the county.
Foodpantries.org: https://www.foodpantries.org/co/ca-los_angeles
Grocery Outlet Store Locator:
https://www.groceryoutlet.com/store-locator?store_location=&store_region=Los+Angeles

Description: The foodpantries.org website offers a list of food banks in the county, while the Grocery Store outlet store locator website offers a list of Grocery Outlet locations in the county.

Features included:
  ● Locations of food banks
  ● Locations of Grocery Outlet stores

Variables:
Food Bank Name
Food Bank Address
Food Bank Zip Code
Food Bank Latitude
Food Bank Longitude
Grocery Store Name
Grocery Store Address
Grocery Store Zip Code
Grocery Store Latitude
Grocery Store Longitude

Data extracting and data cleaning:
  ● Data was obtained by scraping the foodpantries.org and Grocery Outlet websites.
  ● Addresses were converted to coordinates (latitude and longitude) using the geopy package in python

Data storage:
Extracted and cleaned food datasets are stored in Cloud Firestore. To connect with Firestore we generated a private key. Normally you should NEVER store a key in a public GitHub. Therefore, we used Secrets Management in Streamlit sharing to securely connect to private data sources. The sample python code of creating TOML secrets provided. (key-to-toml.py)

Education Equity:

Description: This data was obtained from https://www.greatschools.org/. All schools in LA County can be searched by their LA County zipcode. The number of schools can reflect the distribution and concentration of education around the county. Ratings reflect the evaluation of

each school as well as the diversity and inclusiveness within the school. The teacher-student ratio can reflect the educational resources that each student can receive. Through these, we can understand the educational equality of the whole county.

Features included:
School distribution and density distribution
Multi-tab search based on school type
Multi-tab search based on school grade
List multi-label search results
Mouse suspension shows the proportion of teachers and students
Search for schools by zip code
Table of Postal Code Search Results
Count by school type and grade

Variables:
sch_name: school name
sch_add: school address
zip_code: school zip code
add_dist: school block
sch_rating: school evaluation (based on test scores and school equality ratings)
sch_StdPerTchr: school-teacher-student ratio
sch_type: school type
sch_garde: school grade
longitude: longitude
latitude: latitude

Data extracting and data cleaning:
Obtain the required zipcode from the LA County Open Data website (https://data.lacounty.gov/GIS-Data/ZIP-Codes-and-Postal-Cities/c3xr-3jw2/data). Use the selenium package to scrape data from greatschool's dynamic web page, based on the zipcode obtained earlier. Use the geocoder package to access the Bing map service to obtain the latitude and longitude corresponding to the address. Clean up the data, check whether the newly acquired address matches the previous address, and correct the erroneous data. Insert missing data, according to the principle of proximity.
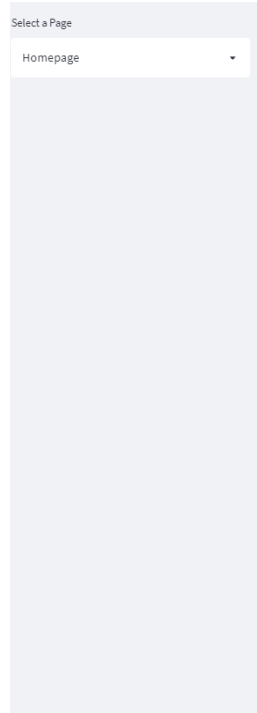
Data storage:
Data is stored locally and uploaded to firebase cloud database.

Screenshots of main functions:

Homepage:
The default page of the application

Select a Page

Homepage ▾

Social Inequality Across LA County

Topic: **This project will examine inequalities in regions of Los Angeles based on crime rates, access to food, and access to quality education.**
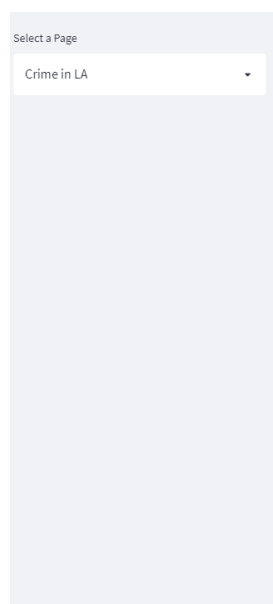
The app is made by *Yerkebulan B., Soumeya K., Zihao H.*

Original Map of Los Angeles Country

"Crime in LA" page
Page related to the crime dataset

Select a Page

Crime in LA ▾

Crime in LA

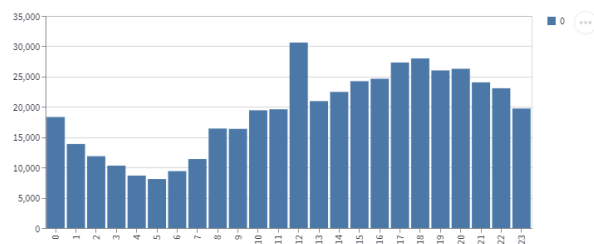Done! (using cache)

Insert a number of ZipCode or type/check "ALL" to see all crimes in LA county

☑ ALL
☐ Show raw data
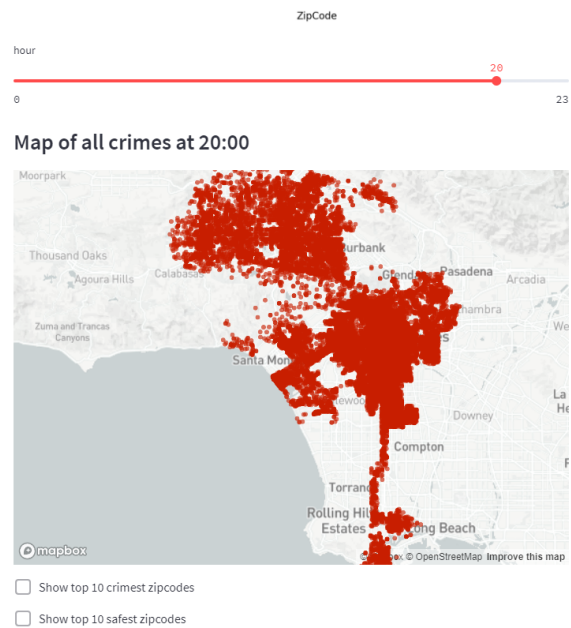
Number of crimes by hour

ZipCode

hour

20

0   23

Map of all crimes at 20:00

☐ Show top 10 crimest zipcodes

☐ Show top 10 safest zipcodes

Select a Page

Crime in LA ▾

Top 10 crimest zipcodes

	zipcode	Count
2	90003	14479
32	90037	12551
24	90028	11145
5	90006	10815
45	90057	9712
33	90038	9336
25	90029	9256
130	91405	7802
48	90062	7723
105	91303	7484

☑ Show top 10 safest zipcodes

Top 10 the safest zipcodes

	zipcode	Count
65	90245	2
138	91506	3
91	90802	3
70	90290	4
97	91105	15
102	91206	17
96	91042	20
93	90831	21
99	91203	40
94	91030	55

☐ Show top 10 crimes

Top 10 crimes

	crm cd desc	Count
130	VEHICLE - STOLEN	50962
4	BATTERY - SIMPLE ASSAULT	36788
20	BURGLARY FROM VEHICLE	30024
126	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)	29952
19	BURGLARY	28500
2	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	27505
114	THEFT PLAIN - PETTY ($950 & UNDER)	24665
70	INTIMATE PARTNER - SIMPLE ASSAULT	24115
112	THEFT OF IDENTITY	22389
110	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	20041

☑ Show top 10 rare crimes

Top 10 rare crimes

	crm cd desc	Count
62	GRAND THEFT / AUTO REPAIR	1
56	FAILURE TO DISPERSE	2
60	FIREARMS EMERGENCY PROTECTIVE ORDER (FIREARMS EPO)	2
107	TELEPHONE PROPERTY - DAMAGE	2
86	PICKPOCKET, ATTEMPT	2
67	INCEST (SEXUAL ACTS BETWEEN BLOOD RELATIVES)	3
61	FIREARMS RESTRAINING ORDER (FIREARMS RO)	3
9	BIGAMY	3
63	GRAND THEFT / INSURANCE FRAUD	4
15	BRIBERY	4

☐ Show number of crimes in context of victim sex

15	BRIBERY	4

☑ Show number of crimes in context of victim sex

Number of crimes by victim sex
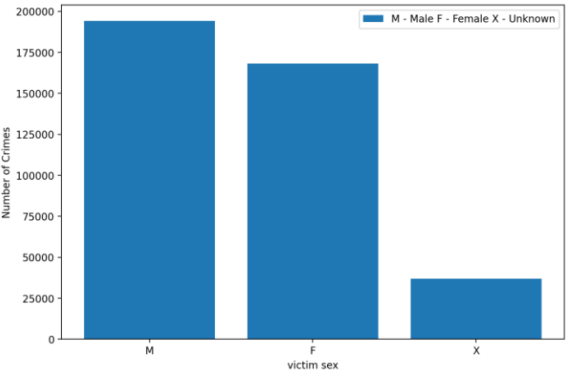
"Food in LA" page
Page related to the food datasets

Map of food banks

Map of discount grocery stores

In addition to food banks and grocery stores, essential food can be obtained by applying for Cal Fresh benefits:

[Cal Fresh Benefits Application and Information](#)
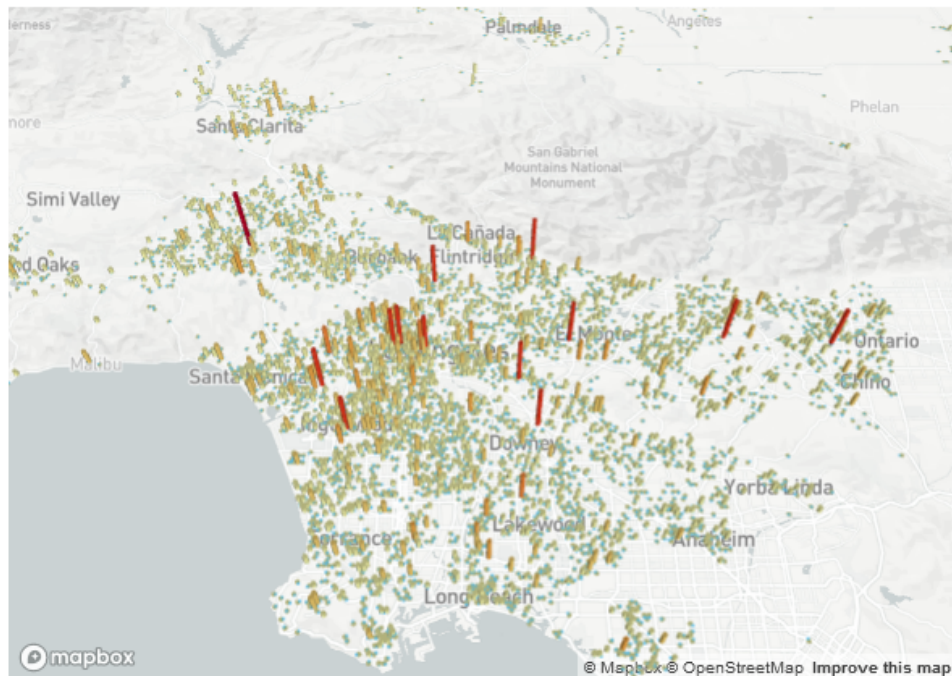
Education Opportunity

Schools

there are 6758 schools in data set

☐ Show Raw Data

Education opportunity Density Map in Los Angeles County:

Scroll to zoom in/out

Search Schools

Which kind of school types would you like to see?

Public district ✕   ⊗ ▾

Which grade would you like to know?

K-6 ✕   ⊗ ▾

☑ Want to hide no rating school?

hidhing no rating shcool

There are 451 schools found based on your preference

Show School detials   +

There are 451 schools found based on your preference

Show School detials —

|	sch_name	sch_rating	sch_StdPerTchr	sch_add
2	Esther Lindstrom Elemen...	6.0000	27:1	5900 Canehill Ave, Lakew
46	Judith F. Baca Arts Acade...	4.0000	23:1	1536 E 89th St, Los Angel
48	Ninety-Ninth Street Elem...	5.0000	24:1	9900 Wadsworth Ave, Los
49	Ninety-Second Street Ele...	4.0000	25:1	9211 Grape St, Los Angele
50	Ninety-Sixth Street Eleme...	3.0000	23:1	1471 E 96th St, Los Angel
101	Richland Avenue Element...	6.0000	23:1	11562 Richland Ave, Los A
170	Columbia Elementary Sc...	3.0000	23:1	2640 E Avenue J, Lancast
174	Eastside Elementary School	3.0000	22:1	6742 E Avenue H, Lancast
177	Enterprise Elementary	3.0000	21:1	3730 E Avenue J, Lancast
200	Tierra Bonita Elementary	5.0000	19:1	44820 27th St E, Lancast

Find Schools in your area

Insert a Zip Code

90001

schools in 90001 area

	sch_name	sch_rating	sch_StdPerTchr	sch_type	sch_grad
2288	Alliance Kory Hunter Middle	4.0000	<NA>	Public charter	6-8
2289	Animo Pat Brown	5.0000	21:1	Public charter	9-12
2290	Charles Drew Middle School	1.0000	21:1	Public district	6-8
2291	Children's Collective Co-Op	<NA>	<NA>	Private	PK
2292	Communication And Tech...	4.0000	23:1	Public district	9-12
2293	Dr. Lawrence H. Moore Ma...	4.0000	24:1	Public district	K-5
2294	Florence Avenue Element...	6.0000	22:1	Public district	K-6
2295	Graham Elementary School	3.0000	25:1	Public district	K-6
2296	Graham Head Start	<NA>	<NA>	Private	PK
2297	Green Design At Diego Riv	5.0000	20:1	Public district	9-12

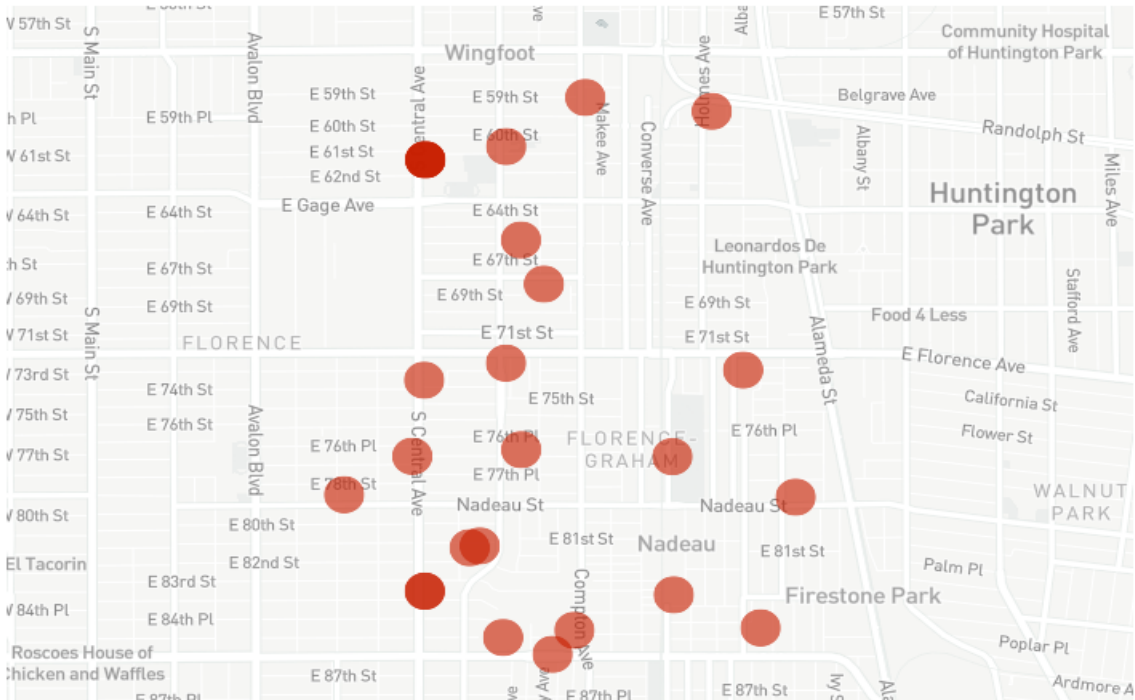## Count by type
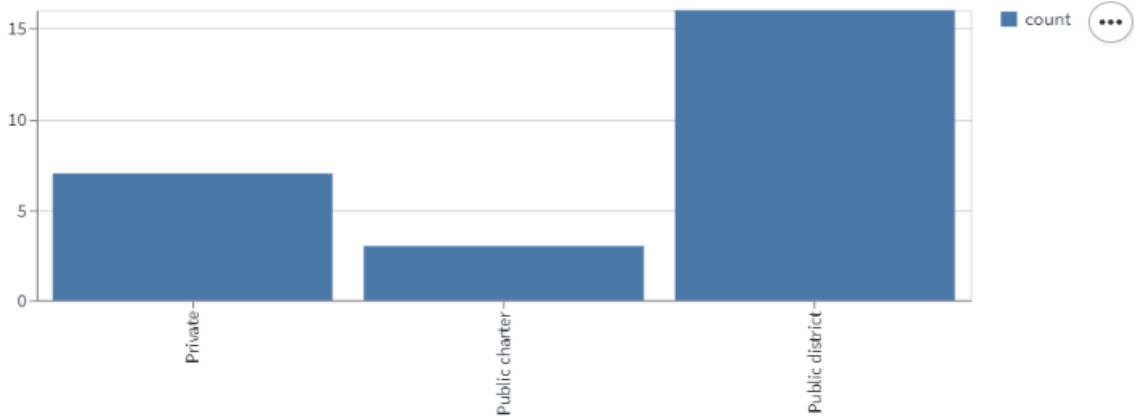
Count by grade
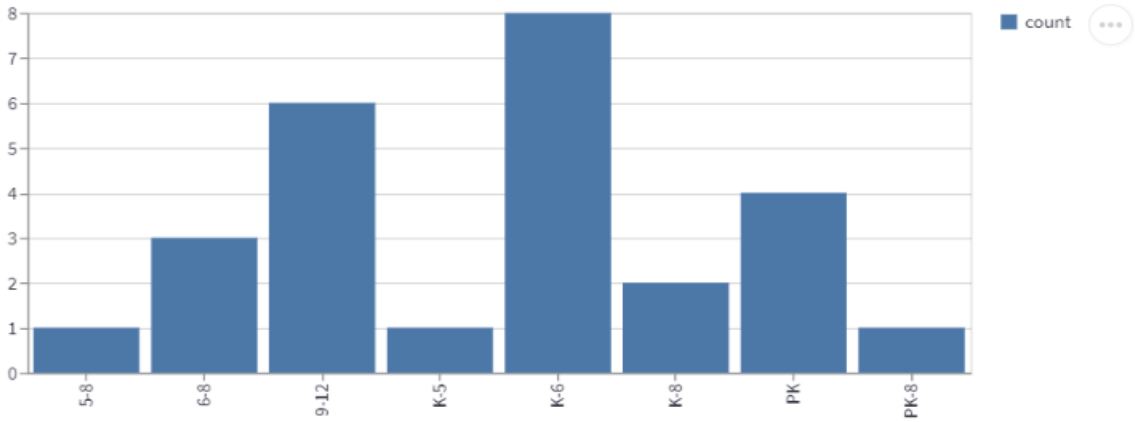
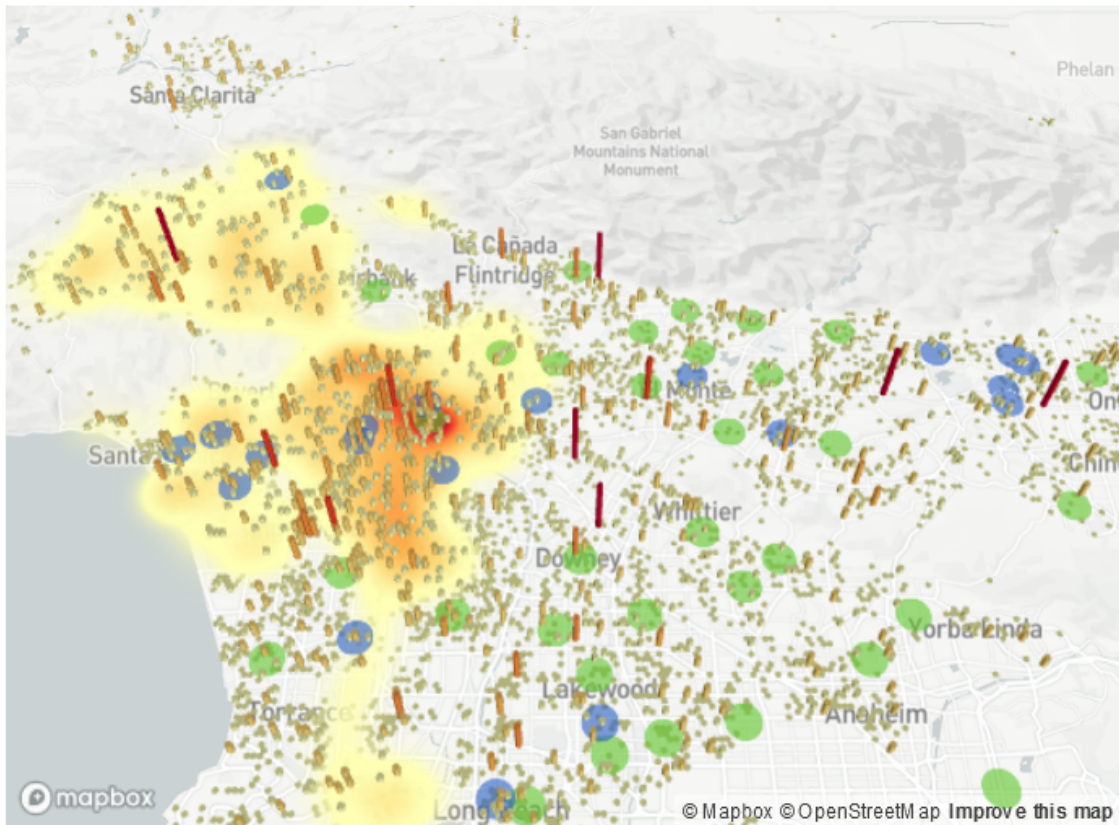## Schools on map

Equity, Opportunity and Risk

Crime, Food, School Map

Scroll to zoom in/out

Bar shows the density of the Schools

Heatmap shows the density of the Crimes

Green Spot shows Food Banks

Blue Spot shows Groceries

Learning Experiences

Finding and acquiring data is our first challenge. In the process of learning and seeking solutions, we have become more clear about our data needs and gradually understand the structure of web pages. We have seen in practice how data cleaning and data quality is a major

challenge in data science. Ensuring that our datasets were properly formatted and had sufficient and quality data was one of the most major hurdles in this project. In addition, we have also seen how powerful large datasets can be in addressing important social challenges. Beyond food, crime, and education, we can also harness similar techniques in this project to dig deeper into related issues such as housing and healthcare.

Studied technologies:

1) Streamlit - an open-source Python library that makes it easy to create and share custom web apps for machine learning and data science. We have studied and used elements such as st.cache, st.map, st.image, st.selectbox, st.slider, st.pyplot, pydeck package, st.vega_lite_chart and etc.
2) Cloud Firestore - a NoSQL document database that lets you easily store, sync, and query data for your mobile and web apps - at global scale. We have studied how to write data to Firestore, how to read from Firestore, and to securely connect a streamlit app with Firestore.
3) Cloud Firebase - NoSQL cloud database to store and synchronize data. Data is synchronized to all clients in real time. The data of the web page will get the data as a json file from firebase.
4) Selenium Web Testing Framework - Cope with dynamic web pages as well as Anti-Spider by simulating real clicks.

Challenges Faced

Crime:
The crime dataset had an inappropriate format of date and time. We had to perform additional data cleaning to convert them into a standard format for further use.
The dataset did not have zipcodes of locations where the crime occurred. We used uszipcode library, which is not working properly in our local machine python environment. Thus, we use Google Colab to get zipcodes.
We were not able to store the dataset in Cloud Databases which we studied in the course, because Cloud Databases, in free plans, have quotas for writing and reading.

Food:
We were not able to obtain the dataset used by USC's Public Exchange for their "Enough to Eat" study, due to data usage agreements. Alternative options online were extremely dynamic and interactive maps which caused a barrier in web scraping. Other websites offered differing information on food access in Los Angeles. We settled on using the list from foodpantries.org which lists some food pantries, food banks and other food charities in Los Angeles County. We also used the Grocery Outlet store locator to scrape Grocery Outlet locations, as this is considered an affordable food option.

Education:
The education dataset we used was extremely large and required a lot of time in cleaning and pre-processing in order to perform analysis on. Dynamic web pages and anti-spider are encountered during the process of obtaining data, and we solve it through Selenium. It is necessary to match other data sets so it is necessary to get latitude. However, large quantities from bing map server requests cause the server to overload, resulting in missing data. Therefore, we added the SLEEP delay and separate the data to multiple batch requests.

Equity:
The combination of three datasets is extremely large. There are some problems with loading and rendering in streamlit and pydeck package. Too many layers of plot makes plot rendering slow and may experience short delay when zooming in and out or it will freeze if there are too many zooming commands input requests in a short period. It also requires a lot of GPU to process. No solution found yet.

Team Members and Responsibilities:

Team Members	Responsibilities
Yerkebulan Bauyrzhanov	Uploaded data to Cloud Firestore Created a secure connection to Cloud Firestore Extracted and cleaned the crime data Created "Homepage", "Crime in LA '' pages in the Streamlit App. Did some analysis related to crime rates for the conclusion.
Zihao Han	Uploaded data to Firebase Created "Education in LA" page Created "Equity, Risk and Opportunity" page Did education and equity related analysis for the conclusion
Soumeya Kerrar	Scraped food bank and Grocery Outlet store websites for dataset Created "Food in LA" page Created "Additional Resources" page Did food-related analysis for the conclusion Created video and in-class demo scripts

Conclusion:

Crime Rates:

1. According to the histogram, we can notice that the peak of crime rates is at 12:00 pm. However, we guess that it is a default time when the time of the crime is unknown. Therefore, we consider the majority of crimes take place in the evening time with the peak at 6 pm.
2. According to the plot, many crimes take place in a range of zipcodes 90000-90080.
3. The most dangerous regions are locations with zipcodes 90003, 90037, 90028 with 14627, 12675, and 11224 crimes respectively from 2020 to the present.
4. The safest regions are locations with zipcodes 90802, 90245, and 91506 with only 3, 3, and 4 crimes respectively from 2020 to the present.
5. Top crime types are stealing vehicles, simple assault of batteries, burglary from vehicles, and vandalism.

6. According to the categorical plot, crimes against males occur more often than against females

Food Access:

1. Food banks and Grocery Outlet stores seem equally distributed across Los Angeles County, however, there are parts of the Eastern part of the county where access to these food resources is more sparse.
2. Access to CalFresh benefits is one resource which surmounts the physical accessibility to food, as it allows members to obtain financial assistance for food at a variety of places like supermarkets, farmer's markets and some restaurants.

Education Access:

1. We can see from school distribution, although there are schools in each city of La County, the most concentrated region is still in La City, especially the Koreatown area. Education resources are largely distributed in developed large cities.
2. However, in the areas that have most schools, the overall rating of these schools is very low.
3. The large number but low -quality schools have further differentiated limited educational resources.
4. Although under the plan of La Unified School District, LA has the 2nd Largest Public School Distrib of the US. The number of private schools is far greater than that of public schools. We still need to increase the number of public schools to make more people have an opportunity to be educated.
5. It is very obvious that the proportion of teachers and students is unbalanced, and most schools are 20-30student Per Teacher. However, some schools even have a proportion of teachers and students up to 90: 1 and they are the Elementary School or K-12 schools. Children do not get enough care in such schools, so they are more likely to be bullied.

Equity and Opportunity and Improvement:

1. Downtown LA and Koreatown area are the places that have the highest population density in the entire LA county, and also the places have significant inequity.
2. Although there are many schools in these two places, the rating of these two areas are very low. It shows the quality of education is not high, and there may be inequality in the schools.
3. From the perspective of food supply, although the distribution is spared, it's still not enough for these two areas. Unbalanced resource allocation will also increase the crime rate.
4. A large number of people brought a high rich-poor gap, there is unbalanced resource distribution in both places, the relatively high crime rate also illustrated this inequity behavior.
5. Thus, we should pay more attention to the public resources of these two areas, providing more resources, and may help to reduce crime rates. While pursuing educational resources, we should also concentrate on resources and provide higher quality schools to avoid limited resource dispersion. In addition, we should also enhance the security of these regions to prevent children from being in danger at school.

Related links:

Los Angeles County has various public services available to address the issues we have explored in this project, as well as other related issues such as housing and public health. This feature is to supplement the visualizations we have provided, so users are aware of how inequalities in the county are being addressed. What we have seen through development of this application is that the scale of inequalities remains very large and many of these programs need to be expanded to address all these different social challenges affecting a population of Los Angeles County's size.