Final Project Proposal

Group 24

CIS 5500 Spring 2024

03/09/2024

1. ***Group Information***

Name	Email	Github
Jiayi Wang	jess1102@seas.upenn.edu	jess1102
Luca Wu	lucawu@seas.upenn.edu	lucawu0725
Ruimin Yin	yrm@seas.upenn.edu	RuiminYin
Yu Feng	yufeng8@seas.upenn.edu	YuFeng888

2. ***Project Description***

The purpose of this project is to create an accessible and user-friendly data platform designed to aggregate safety information for Philadelphia based on the user's latitude and longitude. This platform allows Philadelphia residents and visitors to receive up-to-date and accurate safety updates for their specific location and includes a data sorting function. Users can select the type of safety information they wish to view, such as robberies, gun incidents, and arrest rates in Philadelphia. For the events that happened, the platform can provide detailed information about each incident. With this comprehensive and customizable safety data, citizens and visitors will be better equipped to make informed decisions about their travel plans, significantly improving their overall safety and enjoy peace of mind.

3. ***Datasets Selected***

A total of 5 datasets will be included for this project. 2 of those are major datasets containing more than 100,000 rows of data that is related to the arrest data and crime incident data of Philadelphia. A more detailed preview of the data are included for those two major datasets indicating some potential query opportunities.

1. Major dataset 1: Arrest Data of Philadelphia:

   ‘Arrests Citywide (CSV)’: https://opendataphilly.org/datasets/arrests/

   The Philadelphia Police Department's arrest database serves as a crucial resource for public safety and transparency. It compiles detailed records of arrests made by the police across the city, including information on the nature of the offenses, the identities of the individuals arrested (subject to legal privacy protections), and the locations and dates of

these incidents. By providing access to this data, the database aids in understanding crime patterns, enhancing community awareness, and fostering accountability within the law enforcement process. This tool is invaluable for researchers, policymakers, and residents seeking to improve safety and justice in Philadelphia. This dataset contains 190,456 rows and 5 attributes.

The racial distribution of defendants is predominantly Black (93,588 entries), followed by White (47,417 entries), and Latinx (39,911 entries). Asian defendants are recorded 5,071 times, and there are 4,469 entries labeled as Unknown/Unreported/Other. The distribution of arrest counts shows a right-skewed distribution, indicating that a large number of arrests involve relatively few individuals, with fewer instances involving a larger number of individuals. The histogram was limited to arrest counts up to 50 to focus on the more common range, excluding outliers.

The result of the chi-square test for independence between 'defendant_race' and 'offense_category' yields a Chi-square statistic of approximately 25321.34 with a p-value of 0.0. This indicates a statistically significant association between the defendant's race and the offense category. In other words, the distribution of offenses varies significantly by defendant race, suggesting that race and offense category are not independent of each other.

2. Major dataset 2: Crime Incidents of Philadelphia:
   'Crime Incidents from 2006': https://opendataphilly.org/datasets/crime-incidents/

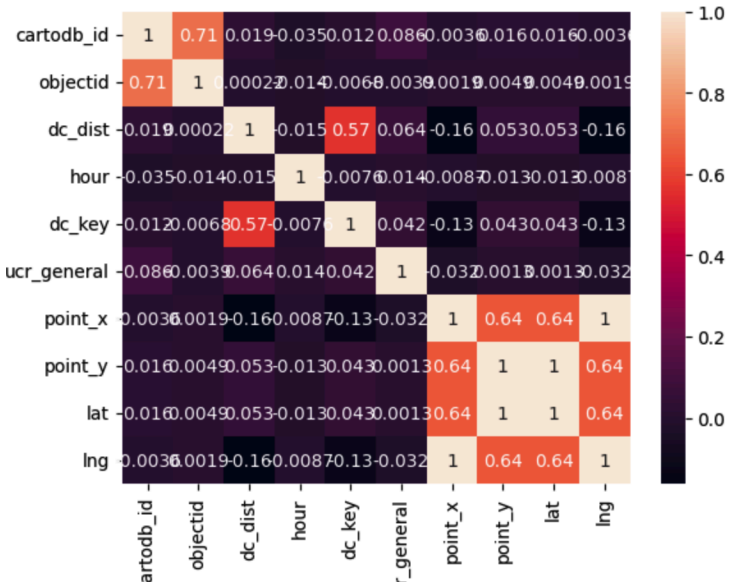
   Crime incidents from the Philadelphia Police Department. Part I crimes include violent offenses such as aggravated assault, rape, arson, among others. Part II crimes include simple assault, prostitution, gambling, fraud, and other non-violent offenses.

   The relevant statistics of the dataset is as the following:

	cartodb_id	objectid	dc_dist	hour	dc_key	ucr_general	point_x	point_
count	1.690170e+05	1.690170e+05	169017.000000	68398.000000	1.690170e+05	169017.000000	161952.000000	161952.00000
mean	2.672247e+06	6.094921e+06	17.534147	13.160867	2.023174e+11	952.305981	-75.148116	39.99541
std	8.968762e+05	3.780515e+06	11.274421	5.647942	1.987874e+07	601.749177	0.061416	0.04631
min	2.000000e+00	9.600000e+01	1.000000	0.000000	1.979351e+11	100.000000	-75.275297	39.87464
25%	3.006752e+06	3.005223e+06	9.000000	10.000000	2.023090e+11	600.000000	-75.184472	39.95985
50%	3.085235e+06	5.940979e+06	16.000000	13.000000	2.023160e+11	700.000000	-75.156301	39.99376
75%	3.127516e+06	8.959895e+06	24.000000	17.000000	2.023241e+11	1100.000000	-75.115646	40.03128
max	3.227229e+06	1.423841e+07	77.000000	23.000000	2.023770e+11	2600.000000	-74.957607	40.13761

By aggregating incident counts over time, you may observe certain trends. For example, there might be more incidents on weekends or during specific months. This could indicate seasonal patterns or the impact of certain events on incident rates.
in offenses (text_general_code) over time can reveal shifts in the types of crimes reported. For example, an increase in theft-related offenses over the years might prompt further investigation into socioeconomic or law enforcement.

Examining the distributions of numerical variables such as hour can reveal when incidents are most likely to be reported. A skewed distribution towards nighttime hours might suggest a higher rate of certain types of incidents occurring during those times.



Graph 1: Correlation between each attribute in the dataset.

3. Supplementary dataset 1: Philadelphia Attorney office data: (https://data.philadao.com/download.html)

   The dataset contains a total of 373,376 rows and 5 columns. Here's a breakdown of its characteristics:

- The `defendant_race` column has 5 unique values, reflecting the diversity of defendants' races.
- There are 26 unique police districts represented, indicating the geographical spread of the data.
- The `offense_category` includes 43 different types of offenses, showing the variety of crimes recorded.

- The `day` column has 4,816 unique values, suggesting the dataset covers a span of approximately 13 years, assuming no gaps in dates.
- The `count` column has 36 unique values, which might indicate the number of times an incident occurred or was recorded in a particular way.

4. Supplementary dataset 2: Shooting Victims of Philadelphia:
   (https://opendataphilly.org/datasets/shooting-victims/)
   City-wide shooting victims, including Police Officer-involved shootings
   This dataset contains 15545 shooting event records. The attributes of this dataset includes:

- 'The_geom' is a numerical indicator of the geometry. It is expressed in the well-known binary(WKB) format.
- 'The_geom_webmercator' is a numerical special field automatically added by carto to speed up rendering of tiles for mapping services. It is the same as the 'the geom' column but projected in Web Mercator (EPSG:3857).
- 'Object_id' is a numerical indicator unique for every record.
- 'Year' is a numerical indicator of the year of occurrence
- 'Dc_key' is a string indicator containing number that identifies each unique event (ref dataset 5: Philadelphia Gun Violences: 'Dc_key')
- 'Code' is a numerical indicator of the code of the event
- 'Date' is a date indicator of the date of occurrence in the form of (month/day/year time)
- 'Time' is a string indicator of the exact specific time of occurrence (hh/mm/second)
- 'Race' is a string indicator with single letter to indicate the race of the offender ('B' or 'W')
- 'Sex' is a string indicator with single letter to indicate the gender of the offender ('M' or 'F')
- 'Age ' is a string indicator of a number indicating the age of the offender
- 'Wound' is a string indicating the wounded part of body during the occurrence
- 'Officer_involved' is a string indicator with single letter to indicate if an officer was involved in the event ('Y' or 'N')
- 'Officer_injured' is a string indicator with single letter to indicate if an officer was injured in the event ('Y' or 'N')
- 'Offender deceased' is a string indicator with single letter to indicate if the offender was deceased in the event ('Y' or 'N')
- 'Location' is a string attribute of the street name of the event
- 'Latino' is a numerical binary indicating if offender is latino
- 'Point_x' and 'point_y' are both numerical values indicating geographical longitude and latitude coordinates of the event respectively

5. Supplementary dataset 3: Philadelphia Gun Violence: https://controller.phila.gov/philadelphia-audits/mapping-gun-violence/#/?year=2024&map=11.00%2F39.98500%2F-75.15000

   The dataset contains 15515 locations of shooting victims. The attributes include:

- 'Dc_key' is a string indicator containing number that identifies each unique event
- 'Race' is a string indicator of the ethnicity of the offender
- 'Sex' is a string indicator of the gender for the offender in the form of (male or female)
- 'Fatal' is a string binary indicator for the outcome of the event in the form of (Fatal or Non-Fatal)
- 'Date' is a string indicator of the date of occurrence in the form of (month/day/year time)
- 'Has court case' is a binary string indicator for if there is associated court cases (Yes/No)
- 'Age' is a numerical indicator of the age of the offender
- 'Street name' is a string attribute of the street name of the event
- 'Block number' is a string attribute of the block number of a street (references 'street name')
- 'Zip code' is a 5-digit numerical number that identifies the zip where the event occurred
- 'Council district' is a numerical number indicating the council district where the event occurred
- 'Police district' is a numerical number indicating the police district where the event occurred
- 'Neighborhood' is a string indicating the neighborhood where the event occurred
- 'House district' is a numerical number indicating the house district where the event occurred
- 'Senate district' is a numerical number indicating the senate district where the event occurred
- 'School catchment' is a string indicating the school located where the event occurred
- 'Lng' is a numerical number indicating the geographical longitudinal location of the event
- 'Lat' is a numerical number indicating the geographical latitudinal location of the event

4. Proposed Queries In Natural Language

- Aggregation with Condition: Calculate the average number of crimes reported per month in the vicinity of each public school in Philadelphia. 'Vicinity' will be defined by latitude and longitude coordinates.

- Join and Aggregation: Identify the top five neighborhoods with the highest arrest rates and compare them to the average number of shooting incidents in those neighborhoods over the last year.

- Complex Query with Subquery: Find all instances where the number of shooting victims in a given zip code exceeds the monthly average across the whole city for the current year.

- Complex Join involving Multiple Datasets: For each district, determine the number of violent crimes that led to arrests, and of those, identify how many involved repeat offenders, using the Attorney office data for offender history.

- Join and Aggregation: Aggregate the total number of gun violence incidents and shooting victims within a half-mile radius of major transit stations and rank these stations by the frequency of incidents.

- Join and Aggregation: Based on empirical data, predict the regions for next month's highest crime rates, focusing on specific areas defined by latitude and longitude coordinates.