How Does Crime Affect Bike-Sharing Trips in San Francisco?

Progress Report

Tsai-Tzu Cheng
Applied Science and Engineering
University of Colorado
Boulder CO USA
tsch3115@colorado.edu

PROBLEM STATEMENT AND MOTIVATION

The team is interested in data mining the user trip data of Bay Wheels in the bay area (including San Francisco), with the San Francisco crime report to discover whether the historical/recent criminal incident locations has taken effects to the biker route decisions, and the period of the occurrence.

In addition to answering these, the team also wants to form analytical theories. For instance, if the team finds a contrast between an immense of sudden vacancies appearing at a certain bike parking location and the high crime report frequency around there, the team will theorize what reasons to cause biker behavior changing when the crime rate is high. The goal is to uncover the strongest and most pronounced correlation or pattern in the data.

LITERATURE SURVEY

The recent research conducted by Rimolo, Silvia examined the correlation between selected common street crimes and potential increases in human activity caused by bike sharing ridership in Boston and Cambridge, MA. Silva hypothesized that Boston and Cambridge neighborhoods with a higher ridership of BSS have lower incidents of robbery, assault, and larceny, and that ridership correlations with robbery, assault, and larceny incidents are on par with, or stronger than, correlations to other demographic factors. Using various statistical and software applications, Silva analyzed BSS ridership against several relevant crime variables, and included other demographic, economic and geographic factors that the FBI considers as recurring factors associated with variation in crime rates.
The results showed a positive correlation between BSS ridership and nearby crimes, indicating that when ridership increased, nearby instances of selected crimes increased as well.

PROPOSED WORK

To begin data collection, the team must first preprocess the two data sets: the San Francisco crime report data set from [3], and the Bay Wheels trip data set from [4]. Combined, our data sets have an immense number of rows. If we wish to find correlations between crime types, occurring time, and bike trips taken, we will need to first clean the data. The first step will be to remove the null/missing values, especially in the Bay Wheels trip data set. There are some crucial works need to be done, for instance, some trip data is missing the start station value, in this case, we can use the longitude and the latitude to identify the location, however, these numerical attributes are not very accurate when displayed in floating points. Discovering correlations requires transforming the data. For the crime data, the team will first reorder the data set in chronological order, then group the data by the type, and location. Later, the trip data will be sorted by the started time, and then grouped by the [start, end] station. With the given time and location from the crime data set, the team will be able to easily compare it with the trip data set. We may even derive a new attribute type that calculates the Euclidean distance divided by the travel time; with the crime data, this attribute may give us some insight, such as finding if the biker took an extra ordinary long trip to avoid some crime spots.

DATA SET

Our analysis will primarily consist of utilizing two data sets. One of which is the crime report data, and the other is the biker trip data (Figure 1). We intend to focus exclusively on San Francisco, California. This ultimately means our data and analysis will only be for the downtown San Francisco area (Figure 2).

Figure 1: May 2024 Bay Wheels trip data

Date	Time	Incident #	Location	District	Category SFPD	Description	Resolution
07-05-2024	00:00	240418554	Mission	Mission	Motor Vehicle Theft	Vehicle, Stolen, Truck	Open or Active
07-05-2024	00:00	240419580	Outer Mission	Ingleside	Larceny Theft	License Plate, Stolen	Open or Active
07-05-2024	00:00	240419245	Bernal Heights	Ingleside	Larceny Theft	License Plate, Stolen	Open or Active

Figure 2: Up-to-date crime statistic report

EVALUATION METHODS

To evaluate our result, we plan on correlating all crime incident data with clusters of historical biker trips. Firstly, our data set provides us with a time, coordinates, and the description of the crime. Using this information, we can plot a map of the crime locations within a certain time. Subsequently, we will mark the biker trip start-end locations to compare with. We can also create a window for all bike stations sorted by the least settled station and sum up all the crimes that fall in that window. With this, we can create a ratio between station idleness and the number of crimes. We can then analyze the behavior displayed between crime and the biker decisions.

TOOLS

For the team to best find correlations between our data sets, the team will use a few different tools to accurately find patterns. To arrange and look at the data, the team will utilize Excel which provides useful data-centric tools to manipulate the data with ease. Aside from performing cleaning and arranging the data, the team will use Excel for some basic analytics. By making use of highlighting certain rows, or placing certain columns next to one another, the team can make some initial predictions about possible patterns.

The team will store the data sets into databases using PostgreSQL Server. By using the SQL server, the team can create specific queries to look at certain patterns. For example, if the team wants to see the major biker routes for a specific date range and compare that with the crime type and location for that date range, the team could easily create a SQL query to review that data, and then takes that data and imports it into Excel for further analysis.

The main tool the team plans on using for data analysis is WEKA. WEKA uses machine learning to help specify patterns within a dataset. WEKA also has a simple user interface that would help identify correlations faster. The team also plans to use python scripts to analyze the data. Due to the immense size of our data sets, utilizing specific python scripts can help easily iterate through all the data to find specific values and patterns quickly.

For visualizing our data, the team plans to use Matplotlib. Matplotlib allows for very easy plotting of data and offers a wide range of different plot styles. Being able to plot the data in many different graph styles could really help us identify outliers, as well as patterns.

MILESTONES

To begin with, the team will need to clean up the data. The first milestone will be data cleaning and pruning. It may also be necessary for the team to fill in crucial missing data with matching data found from other sources. After that, it is necessary to arrange and sort the data in a way to easily compare each with the others. This will make identifying patterns, locating specific time periods in the data set, and isolating important segments of the data much more efficient and smoother.

Once the above is complete, the team will analysis the data. It will begin by looking for general patterns over a large time period. In doing this the team will see which years or decades had the most crime overall, and we will be able to line that up with the bike trip data to see if there is anything in common. From there, the investigation can go deeper, using smaller time scales to find more precise patterns.

The team will document our findings as the progress goes along, and keep the repository updated on GitHub. The next milestone will be to find a solid pattern in our data that we can graph or document and submit to GitHub. The team will continue to adjust and refine the in-progress work as the team understands the data better and how it relates, but for now, the next milestone will be to tackle the problem statement questions. The refinement continues until the team makes confident conclusions to these questions.

Finally, the team will need to create our presentation to show the findings to the class. The final milestone, for now, will be a clear and concise project presentation that elegantly explains the team's work to peers in a way that will intrigue and inform them. By this point, the team hopes to show a solid

understanding of the discovery and how data relates and hopes it can thoroughly answer any questions the classmates may have.

Milestones Completed

Most of the work we have completed up to this point has mainly been developing the Python script that is able to not only clean up and organize the data but also to derive useful new attributes explicitly through the third-party APIs for analysis. While this initial work has not resulted in direct visual results, it will hopefully allow us to more easily complete the difficult and time intensive analysis portion later.

The initial crime data was messy and had some redundant data and other issues, for example, the datetime attribute for each row is split into date and time, with the date value being "2024-06-01 00:00:00" and the time value being 2024-06-01 01:03:00. In this case, it's clear that we should merge two attributes into one. There were hundreds of thousands of cells in the crime data set belonging to columns that were not helpful to find the correlation with the bike trip data, so those data points had to be removed.

The key accomplishment up to this milestone is unifying the geo location value which can link the incidents on the crime data set to the trips on the bike trip data set by integrating the Google Geocoding API into our Python script. Originally, the geo values used on the crime data set were shown as some vague address, while the geo value used on the trip data set were rendered in not only street addresses but also latitude/longitude up to 15 decimal places. Geocoding is well-suited to resolve this inconsistency and to truncate the unnecessary accuracy.

Milestones Todo

By now the Python clean up scripts are completed and tested against small portion of raw data sets. The team needs to catch up the progress to clean up the whole data sets.

The team also needs to dig in our data analysis and related work more. With the data cleaned and grouped, we will be able to do the analysis much more efficiently. This should make the analysis portion much quicker, especially with the derived data sets we have, to easily reference what time of day and what location in San Francisco we will be focusing on for the interesting bike trip patterns.

While we have begun to look at the correlations between bike trip locations, crime locations, and time, we still wish to look for how other attributes play factors in correlations, the crime category for example. By looking at the geographic coordinates and incident start time of the crime, we can begin to analyze correlations between crime and bike trips. To make this easier we an additional data set csv storing frequent street-coordinates key value pairs to cluster location based on proximity, and then compare it to a cluster of crime data based on crime type. To cluster the data, we plan to either partition it through the k-medoids or model-based approach. Once the data is put into clusters, we should be able to find patterns between the two by looking at things like the frequency of certain crimes in a specific area.

RESULTS SO FAR

Most of the results thus far show a very basic look at our data and analysis.

	started_at	start_lat	start_lng	category
1	2024-06-01 00:00:00	37.7597412	-122.4011391	Missing Person
2	2024-06-01 00:00:00	37.7936658	-122.3964729	Motor Vehicle Theft
3	2024-06-01 00:00:00	37.7936658	-122.3964729	Motor Vehicle Theft
4	2024-06-01 00:00:00	37.7808535	-122.4664321	Fraud
5	2024-06-01 00:00:00	37.7936658	-122.3964729	Burglary
6	2024-06-01 00:05:00	37.7936658	-122.3964729	Suspicious Occ
7	2024-06-01 00:11:00	37.7936658	-122.3964729	Fire Report
8	2024-06-01 00:15:00	37.7936658	-122.3964729	Malicious Mischief
9	2024-06-01 00:54:00	37.7389229	-122.4943071	Weapons Carrying Etc
10	2024-06-01 00:54:00	37.7389229	-122.4943071	Stolen Property

Figure 3: Sanitized crime data ready to analyze

	started_at	ended_at	start_lat	start_lng	end_lat	end_lng
1	2024-06-04 07:47:22	2024-06-04 08:15:45	37.7810483	-122.4653783	37.7944659	-122.3947991
2	2024-06-04 13:24:41	2024-06-04 13:28:02	37.7894171	-122.401109	37.7944659	-122.3947991
3	2024-06-06 09:06:34	2024-06-06 09:12:01	37.7894171	-122.401109	37.7807183	-122.3950843
4	2024-06-10 17:44:01	2024-06-10 17:48:29	37.7810483	-122.4653783	37.7821146	-122.4827523
5	2024-06-11 15:52:00	2024-06-11 16:12:14	37.8043547	-122.4161315	37.7807183	-122.3950843
6	2024-06-13 15:55:40	2024-06-13 16:06:59	37.769006	-122.3863379	37.787552	-122.4066183
7	2024-06-23 16:36:36	2024-06-23 16:57:29	37.7639631	-122.4157476	37.7836262	-122.3898681
8	2024-06-25 09:10:37	2024-06-25 09:16:15	37.7894171	-122.401109	37.7836262	-122.3898681
9	2024-06-27 15:36:19	2024-06-27 15:48:22	37.769006	-122.3863379	37.7944659	-122.3947991
10	2024-06-30 10:42:20	2024-06-30 10:51:09	37.769006	-122.3863379	37.7807183	-122.3950843

Figure 4: Sanitized bike trip data ready to analyze

	station_name	lat	lng
1	14th Ave at Geary Blvd	37.7810483	−122.4653783
2	Market St at Steuart...	37.7944659	−122.3947991
3	23rd Ave at Clement...	37.7821146	−122.4827523
4	In Chan Kaajal Park	37.7639631	−122.4157476
5	Delancey St at Brann...	37.7836262	−122.3898681
6	Potrero Hill	37.7597412	−122.4011391
7	Financial District/So...	37.7936658	−122.3964729
8	Inner Richmond	37.7808535	−122.4664321
9	Sunset/Parkside	37.7389229	−122.4943071
10	Montgomery St BART	37.7894171	−122.401109
11	South Park St at 3rd...	37.7807183	−122.3950843
12	Francisco St at Colu...	37.8043547	−122.4161315
13	Terry Francois Blvd ...	37.769006	−122.3863379
14	Union Square	37.787552	−122.4066183

Figure 5: Additional street-latitude-longitude lookup map to assist clustering

REFERENCES

[1] Han, Jiawei, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Elsevier Science, Burlington, 2011.

[2] Bicycle Share Systems: A Predictor of Crime? https://dash.harvard.edu/handle/1/37374915

[3] San Francisco Crime Report https://www.civichub.us/ca/san-francisco/gov/police-department/crime-data

[4] Bay Wheels Data https://s3.amazonaws.com/baywheels-data/index.html