Final Report

Team 008 – Krishna Kumar, Shilpi Jain, Christopher Manley, Kalvin Edwards, Sumit Machwe, Aditya Tadakaluru

1. Introduction – Motivation

Neighbourhood safety is crucial for any new homeowner or renter [1]. Many a times, an area that is considered safe may or may not be safe for an individual or a group of individuals [2]. Unlike general safety ratings, we want to personalize the safety ratings based on the user's profile i.e., find the areas that are safe or unsafe for an individual associated with specific user profile. The specific type of crime, and specific victims of those crimes, are likely to differ across different locations and contexts [3] for e.g., economic conditions which appears to impact safety rating of a particular zip code [4]. Because of its high crime rates and availability of large crime dataset, the city of Chicago is chosen as the region of interest.

2. Problem definition

Currently used apps such as Trulia, Zillow, Crime grade and City Data provide overall general safety rating of a zip code without any personalization i.e., they do not take the factors associated with given user's profile into account. These websites use factors such as school ratings, dwelling attributes and comparable etc. that are limited in scope when it comes to personalization.

3. Survey

The novelty of this approach is user centric personalized safety rating. Theories such as "clear victimology theory" is used in creating user profiles and identifying associated factors to evaluate the safety rating of a zip code. According to this theory the individual demographics are key to the probability of being a victim where factors such as age, sex, ethnicity, race, lifestyle [5][6], affluency can impact the type and probability of vulnerability [7]. Our approach emphasizes that Crime rate is not empirically related to race. However, selection bias and non-representative data contaminates crime zoning [8].

Our solution is beneficial to anyone considering moving to Chicago [9], current residents of Chicago, city officials and public policy makers, law enforcement agencies, insurance agencies [10], education boards and prospective businesses planning on investments in the area [11] can benefit from this application. Home buying and renting decisions within a metropolitan area are largely based on the perceived attractiveness of individual neighborhoods [12]. Having the ability to give more insight on any type of crime for a particular location is invaluable information for prospective or current residents. For e.g., to be successful, street robbers must attack at the right time at the right place [13].

This research will not only help in making informed decisions on what areas of Chicago to move into, but also help law enforcement officers and will help significantly reduce the crime [14]. Law enforcement and governance bodies can measure policy impact on communities [15]. By suggesting alternatives options in neighborhood choice, our application will complement the nation's housing organizations through increasing racial and economic diversity within neighborhoods [16]. There is always a risk that City planners do not take affirmative action and App Users simple avoid hotspot zip codes. This may result in further deterioration of such neighbourhoods [17]. But Real estate investors can influence government/law enforcement to reduce crime rates and increase the price of properties by uplifting the neighbourhoods [18].

4. Proposed method

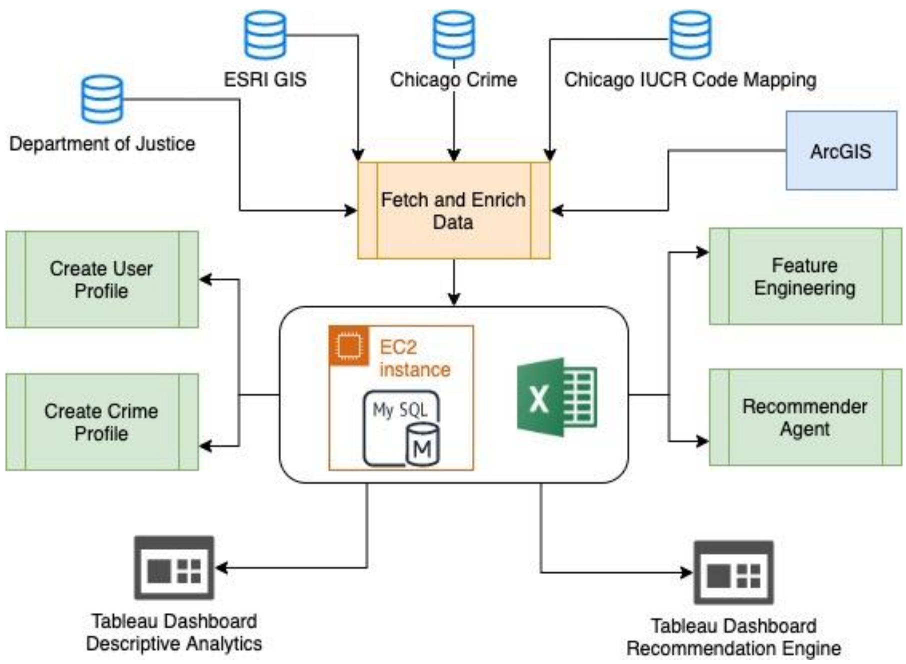

Fig 1: Solution Architecture

Proposed solution has two main modules 1. GUI and visualization 2. Modelling and Analysis. Tableau dashboards are chosen as the main platform for implementing GUI and visualization while a combination of Tableau, MySQL, R and Python are used for the modelling and analysis component. Below is a brief overview of the solution architecture. Our tool can generate sophisticated user profiles of prospective renters, real estate home buyers via feature engineering large datasets. This is unique to our application and is not available with any existing market tools (Zillow etc.)

4.1 GUI and Visualization

Personalized safety rating app is a map-centric tableau dashboard and consists of three main components explained in the order aligning with the user's workflow as the app is accessed.

Dropdown control: User is presented with drop-drown controls that are populated with user profiles that are selected and curated in the modelling and analysis module. This is the first step that user would perform when they access the dashboard. The profiles are built to reflect a person's current state in life both personal and professional for e.g., Senior citizen, Married with Children, Single and working professional, Night shift worker etc. Each profiles is mapped to a combination of attributes that are used in modelling the personalization aspect.

Map control: The map control is driven by the selection of a profile in the drop-down control. The map has two options – a choropleth map of zip code boundaries and a heat map with Chicago city boundary polygon as the outer boundary. Map reflects the safety ratings associated with selected profile, enabling user with easy identification of most suitable zip codes or areas (in case of heat map).

Floating panel: The floating panel control is used to show the top 10 safest and top 10 unsafe zip codes for the user profile chosen in the drop-down control. The user may further expand this top 10 list by clicking on a expand to access the complete list of zip codes and sort by their associated safety ratings.

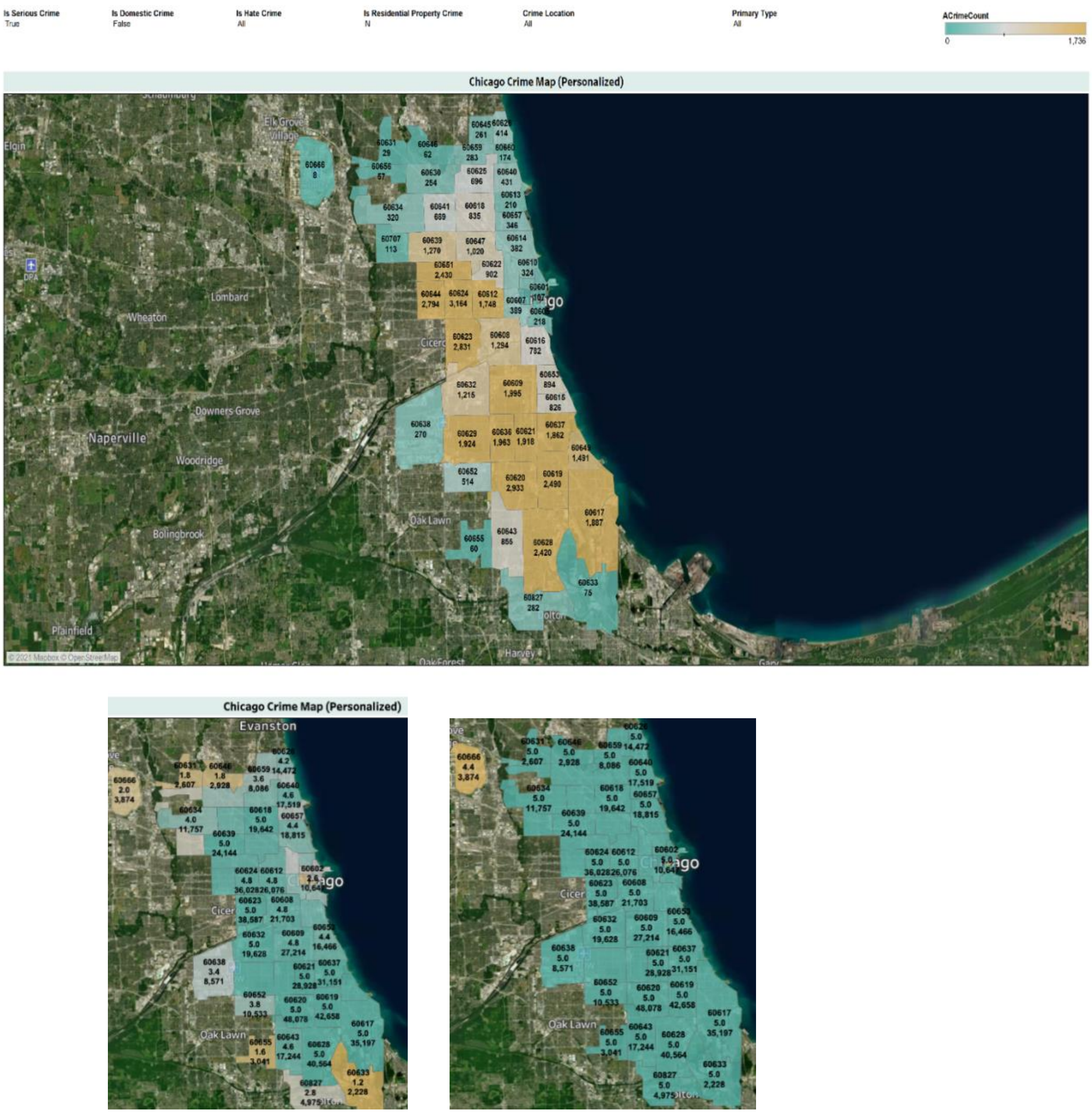

Fig 2: Tableau dashboard showing zip code ratings for different profile settings

4.2 Modelling and Analysis

We have used a combination of Tableau, MySQL and Python are used for the modelling and analysis component. There are four main components in this module each explained below and how they support the visualization functionality in the dashboard.

Zip code rating using Crime Type: Chicago dataset is spatially joined with Chicago zip code layer to extract a zip code for each crime record in the dataset. The zip codes are then mapped to a matrix of simulated user profiles and crime profile by the recommender engine discussed below using the probability distribution for being victimized in a zip code derived from the summary statistics calculated from historical crime data.

Feature Engineering: For successfully modelling personalized crime zones, we need to correlate crime with the time of the day. We create “isDay” and “hour” derived features. Crimes are related to location of Crimes and hence Crime location is enriched to correlate location group to crime occurrence. Not all crimes are serious and based on IUCR codes (Illinois Uniform Crime Reporting codes) - it is segregated between serious vs. non-serious crimes. Crime data is joined with hate crime database to find hate crimes (racial, anti-gay, ant-Jewish etc.). The city of Chicago is chosen as the region of interest because of its high crime rates and availability of large crime dataset. The original dataset downloaded from Chicago data portal contains 1.67 million crime records for the past five years

Leveraged Database SQL query, python/pandas, and geopy to create below features:

i. `IsDay` and `hour-of-the-day`: Correlation of Crime with the time of day
ii. `IsDomestic` : Is it a domestic crime?
iii. `Latitude`, `Longitude`, `Location`, `ZipCode`, `Location-Group`: Location coordinates to correlate crimes with location.
iv. `IsResidentialProperty`: Is crime related to residential property?
v. `IsSeriousCrime` : Is it a serious crime based on IUCR code?
vi. `IsHateCrime`, `HateCrime attributes`: Is it Hate crime?

Mapping Profile susceptibility to a Crime Type: Probabilities are calculated for each demographic attribute (e.g., age, income) against each crime type using the year 2019 from the Bureau of Justice website and stored in a reference table. We found that there were seven crime types that we could use that had a probability distribution for different classes in each attribute of demographic profile available in survey. We made certain assumptions that each attribute is independent and the crime susceptibility of a user profile is a multiple of probability of each attribute.

The probability distribution of crime susceptibility to the 7 crime types () is used as an input by the recommendation engine to generate crimes susceptibility for a finite set of user profiles.

Optimization of safety rating between profiles: Because of the large size of Chicago crimes dataset, the number of factors and the number of profiles that can be built from the combination of these factors are very high in number, making the process of curating the profiles by hand an extremely tedious and time-consuming task. Our approach to mitigate this inefficiency is to reuse the safety ratings generated for one profile to other similar profiles by implementing a user profile recommender system using Cosine Similarity. Using recommender system will allow us to gain both computation and time efficiencies.

Recommender system allows us to connect users and to assist with this process. A user profile simulator is built that will help establish this initial connection. Since the recommender system is an unsupervised learning model and is based on simulated user data we do not have a sufficient way for validating model performance in the initial phase of the application. This will be done through an iterative user feedback approach in future iterations.

5. Experiments/ Evaluation

Bureau of Justice statistics publishes a publicly available, standardized survey database across the US population. Our initial approach was to use this database for feature engineering by regressing each crime type as factor of demographics. During this process we found that the dataset required substantial amount of data engineering. We shifted our strategy to using summary statistics from the database as explained in our solution approach section.

Our key results are as follows:

1. Our recommendation agent provides zip code ratings in the dataset on a scale of 1 - 5 based on average rating given by other users who are deemed similar
2. For a user without assigned profile, Tableau dashboard will identify crime hot spots across zip codes. Based on descriptive analytics, we conclude that:
    - 80% of zip codes have serious crime. Less than 50% of crimes occur around Residential Properties.
    - Hate crimes and Financial Crimes (near ATM and Banks) are sparse and exhibit no recognizable pattern
    - Assault and Robbery are the most predominant crime types
3. Not all zip codes of Chicago are equal in terms of crime distribution, underscoring the fact that user susceptibility depend on personal attributes like demographics, nature of work and lifestyle preferences
4. It is possible to assign a susceptibility score to a user profile at crime type and zip code level. Our implementation calculates probabilities across available demographic features for each individual crime type and derives a combined probability (product) for each user profile and each crime type.

6. Conclusions and discussion

The main assumption (supported by literature review) since the inception is that the safety rating is a factor of demographic profile. We were able to objectively validate this assumption and the same can be observed on the map through noticeable differences in zip code symbology for both general rating of a zip code (Fig 2. Picture on top) and the personalized rating for two different user profiles (Fig 2. Smaller pictures in the bottom).

Our core product and processes demonstrably work for available data. With location specific and other adaptations, the process can be replicated to other geographical areas such as cities, counties, or states.

Some ideas around future work are as follows–

Integration with real estate solutions: By extending our standalone application into a cloud-based web application, it will allow for deep linking from a real estate website like Zillow etc. and serve as a good supplementary tool for users who have already shortlisted the houses and wants to compare the respective zip codes based on personalized safety.

Further personalisation based on lifestyle/demographic info: Create an interactive webpage (chatbot etc.) that can ask the users demographic and lifestyle questions, then dynamically map the user inputs to a profile with the help of recommendation engine.

Further personalization on tableau: The map functionality can be further extended to give user the ability to unselect some hotspots from a zip code based on their lifestyle choices to recalculate the crime susceptibility based on their modified preferences.

7. Team effort - distribution

All team members contribute a similar amount of effort in each phase.

Role	Team Members
Data Sourcing	sjain485, atadakaluru3, kedwards82, cmanley34, kkumar80, smachwe3
Feature Engineering	sjain485, atadakaluru3, kedwards82, cmanley34, kkumar80, smachwe3
Modelling	sjain485, atadakaluru3, kedwards82, cmanley34, kkumar80, smachwe3
Testing and Validation	sjain485, atadakaluru3, kedwards82, cmanley34, kkumar80, smachwe3
User Interface	sjain485, atadakaluru3, kedwards82, cmanley34, kkumar80, smachwe3
Documentation	sjain485, atadakaluru3, kedwards82, cmanley34, kkumar80, smachwe3

References

1. Samarin, Mikhail and Sharma, Madhuri (2020) "A Neighborhood-level Perspective of Real Estate Determinants in Three U.S. Cities," International Journal of Geospatial and Environmental Research: Vol. 7 : No. 3 , Article 2.

2. Madero-Hernandez, A. (2021). Lifestyle Exposure Theory of Victimization. In The Encyclopedia of Women and Crime (eds F.P. Bernat and K. Frailing). https://doi.org/10.1002/9781118929803.ewac0334

3. Connealy, N.T. Can we Trust Crime Predictors and Crime Categories? Expansions on the Potential Problem of Generalization. Appl. Spatial Analysis 13, 669–692 (2020). https://doi.org/10.1007/s12061-019-09323-5
4. Jeff A. Gruenewald & William Alex Pridemore, Stability and Change in Homicide Victim, Offender, and Event Characteristics in Chicago, 1900 and 2000, 13 HOMICIDE Stud. 355 (2009).

5. Bunch, Jackson, Jody Clay-Warner, and Man-Kit Lei. “Demographic Characteristics and Victimization Risk: Testing the Mediating Effects of Routine Activities.” Crime & Delinquency 61, no. 9 (November 2015): 1181–1205. https://doi.org/10.1177/0011128712466932.

6. Deng L. Comparing the Effects of Housing Vouchers and Low-Income Housing Tax Credits on Neighborhood Integration and School Quality. Journal of Planning Education and Research. 2007;27(1):20-35. doi:10.1177/0739456X07301467

7. Wolfgang, Marvin E., and Simon I. Singer. "Victim categories of crime." J. Crim. L. & Criminology 69 (1978): 379. https://scholarlycommons.law.northwestern.edu/jclc/vol69/iss3/8

8. Alex R. Piquero and Robert W. Brame Assessing the Race–Crime and Ethnicity–Crime Relationship in a Sample of Serious Adolescent Delinquents PMID: 19946564 doi: 10.1177/0011128707307219 https://www.ncbi.nlm.nih.gov/pmc/article
9. Tita, G.E., Petras, T.L. & Greenbaum, R.T. Crime and Residential Choice: A Neighborhood Level Analysis of the Impact of Crime on Housing Prices. J Quant Criminol 22, 299 (2006). https://doi.org/10.1007/s10940-006-9013-z

10. Hefei Wen, Jason M. Hockenberry, Janet R. Cummings: The effect of Medicaid expansion on crime reduction: Evidence from HIFA-waiver expansions. Journal of Public Economics, ISSN 0047-2727, https://doi.org/10.1016/j.jpubeco.2017

11. Greenbaum RT, Tita GE. The Impact of Violence Surges on Neighbourhood Business Activity. Urban Studies. 2004;41(13):2495-2514. doi:10.1080/0042098042000294538
12. Petras, Tricia. Measuring the effects of perceptions of crime on neighborhood quality and housing markets. 2007. Ohio State University, Doctoral dissertation. OhioLINK Electronic Theses and Dissertations Center, http://rave.ohiolink.edu/etdc/view?acc_num=osu1180319167.

13. Bernasco, Wim, et al. “Do Street Robbery Location Choices Vary Over Time of Day or Day of Week? A Test in Chicago.” The Journal of Research in Crime and Delinquency, vol. 54, no. 2, SAGE Publications, 2017, pp. 244–75, https://doi.org/10.1177/0022427816680681

14. Jain, Vanita, et al. “Crime Pattern Recognition in Chicago City Using Hadoop Multinode Cluster.” Journal of Information & Optimization Sciences, vol. 40, no. 2, Taylor & Francis, 2019, pp. 587–601, doi:10.1080/02522667.2019.15

15. Cho, Y.H. A multiple regression model for the measurement of the public policy impact on big city crime. Policy Sci 3, 435–455 (1972). https://doi.org/10.1007/BF01405347

16. Mukherjee, Anupam, et al. “Chicago Crime Data Analysis Using PIG in Hadoop.” 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), IEEE, 2018, pp. 242–47, doi:10.1

17. Morenoff, Jeffrey D., and Robert J. Sampson. “Violent Crime and the Spatial Dynamics of Neighborhood Transition: Chicago, 1970-1990.” Social Forces 76, no. 1 (1997): 31–64. https://doi.org/10.2307/2580317.

18. Matthew D. Baird, Heather Schwartz, Gerald P. Hunter, Tiffany L. Gary-Webb, Bonnie Ghosh-Dastidar, Tamara Dubowitz & Wendy M. Troxel (2020) Does Large-Scale Neighborhood Reinvestment Work? https://doi.org/10.1080/10511482.2019.1655468