DME Integrative Assignment

Angelo Barisano; 508903

September 16th, 2022

1 Task 1: Plan & Explore: DEfine Questions for project and familiarize with data

1.1 Overall nature of the Data

Data drives security. The wide spread adoption of data information systems to recognize crime-hotspots patterns to increase policing efficiency and protect people. Thus, creating effective information systems for crime prevention is at the center of policy maker and executive branches of governments. This trend has led to the Chicago PD reaching out to this Data Management Team with the request of creating a Data Management System for crime-data in Chicago.

To this end, the Chicago PD and this Data Management Unit is interested on setting the foundation for a FAIR database; i.e. a Data Base ment to flexible and easy in use of slicing and dicing data to find answers and, contemporaneously, make them available.

1.2 research question

The overarching goal of this project is to set the foundations to funnel the existing crime-data into a data model to improve its readability (analytics), accessibility (logical structure & metric-interpretability), and further use in feeding systems for advanced predictive crime prevention.

In an iterative (agile) developmentcylce the following questions have been set up: Chicargo is known for its high homocide rate; this leads to policy makers focusing on this issue the most.

Upon conducting an initial Exploratory Data Analysis (EDA), four major categories could be identified in the existing data:

- Location
- Beat or Police unit; assigned for a year
- Crime Type
- Arrest(Yes/No)
- Time

Based on the data types, the EDA, and the aforementioned goal of creating a flexible DB, the following research questions guide the creation of the database increaing in complexity:

Hypothesis 1 (H1) *How do certain crime types (homicide) patterns distibute by seasonality patterns, by location, by time of day?*

Hypothesis 2 (H2) *Do crime patterns persist? Do homicide patterns persist in eg a certain district; do certain crimtypes persist wrt. location over time ?*

Hypothesis 3 (H3) *Effective allocation of policing resources to prevent homicides: In order to prevent homocides; which "beats" are the most prevalent among homcides? During which time of day (for effective allocation of policing resources)? What type of crimes necessitate the repeated Loging of information? → these cases might need specialists*

Hypothesis 4 (H4) *beat unit allocation; on a yearly basis, which beat units have the most success? Triangulate on a yearly basis which beats had reduced crime in major cateogreis in addition to how this relates to arrests etc. Try to run a regression on this: crime type predicted by arrest on yearly basis as beats are changed on a yearly basis (so new police officers in a beat every year)*

2 Design and Organize: Identify and collect the necessary requirements for the databased based on the questions, data and information needs

In order to address these questions, the relevant information must be broken down into its underlying entities for easier handling.

2.1 What entities of importantce to design this database?

Part 1 identified four broad data categories that are of relevance in this endeavour. These categories are the general labels which are defined by different degrees of entities/ complexities. This DB is not only designed for the analysis of the data at hand, but also as a basis for a transactional system to be integrated in a larger framework. Subsequently, "slicing & dicing" tasks are dependent on the researcher and should not impact future users of the DB.[1] As such, the DB is created for general use; Variables are created in views or in scripting languages.

To start, the location category is defined in decreasing order of complexity by coordinates (lat, lon), beat or sub-district, block, and police district. One district contains multiple beat and blocks. Thus, this category will be centered by District connected to Block, Beat, and Lat/Lon. **The reason why Blocks are not placed below Beat hierachically is that there might be some overlap between different Beats patroling the same block. If it turns out that this is not the case, then Blocks will be the hirachically the lowest component below beat.** Finally, Lat/Lon will be connected to District specifically, independent of Beat. This is because this information is technically independent from all other location information. However, the choice to include this category hierachically directly below District was made based for the logical structuring of location data; in a sense: where would you look if you search for Lat/Lon. Additionally, including this relation does not increase the complexity of the querries too much due to the connection of District to the center node Case, which we will discuss next. The center node of this DB was chosen to be Case. This is because the DB is designed explicitly to only allow recorded crime data, which will improve on the overall data quality; thus, information such as projections or human entered data is minimized as much as possible, so much so that the DB only represents administrative data.[2] The Location category described above is connected to this center node. Every case must have certain location information defined **(depending on crime;**

[1] This would be an example of setting up tables just for temporary use and then be forgotten or even worse reused for a different purpose: I do not yet know what a future researcher will do. But what I know is how to enable researchers the access to a normed form of data.

[2] Synonym for very clean data from a puplic office

This will be further elaborated upon in part 4 & 5 when we discuss relation restruictions. This way, every data entry starts from creating a Case instance, which then can choose the correct Location data. Beyond Location, every Case requires a crime to be a valid instance. Any crime has a IUCR, which is a code for a combination of a Primary Category of crime, such as murder, and a Secondary Category of a crime, such as first degree murder. Thus, Crime data is connected to Case via IUCR, and IUCR is a code for a primary and secondary relation, defining the crime in question. This way this category is in 3rd NF by design. One might suggest in this context, that IUCR might not be brought up to 3rd NF, as new crimes are rarely added in the penal code and IUCR only contains a total of **343** instances, while the secondary description entails **468** and primary description. The rason to bring this category to the 3rd NF is based on 1) the ease of subgroup analysis in one (e.g.) primary category, disregarding any secondary category; i.e. murder is murder after all. Additionally, 2) explicitness and ease of understanding superceeds ease of use in this case. If a new person to the entire Crime sector simply wants to know what penal codes there are, a subanalysis is easier this way. Finally, but not really important, 3) new secondary descriptions would be easier to be added to primary categories. Furthermore, case specific information will be placed in independent relations, such as Arrests or Location Types. While it might appear that IUCR and Arrest are related, this is technically independent of the Crime description; not every crime leads to a clear cut arrest and is, thus, not mandateted by an IUCR itself. Moreover, Location Type does not describe location specific circumstances pertaining to a Block, Beat, or District but rather the circumstance of how the crime happened. **This also means, that the name of the Location Type will be changed in the project so to remove any confusion.** Finally, any Case requires a Datetime object. This information is placed in the Case instance itself, as the log information is one of the most crucial pieces of information. While some crimes might happen at the same time, normalizing this fact does not make sense with Datetime objects. Views are considerably better suited for "slicing and dicing" the time component than normalizing time, as this project does not have the goal of preimposeing an angle of analysis regarding time.[3] **ARREST SHOULD BE ALSO INCLUDED IN CASE AS A BOOL OR NOT!!**

2.2 WHAT VARIABLES SHOULD BE INCLUDED???

I DO NOT UNDERSTAND THIS PART? ALL INFORMATION IS SOMEWHAT RELEVANT; MAYBE NO BLOCK INFORMATION; BUT MAYBE IT IS RELEVANT. ADDITIONALLY, for task 3, there are no MANY TO MANY relationships to begin with! So what do they want in aprt 3 from me? WHAT DO YOU MEAN WITH VARIABLES HERE?

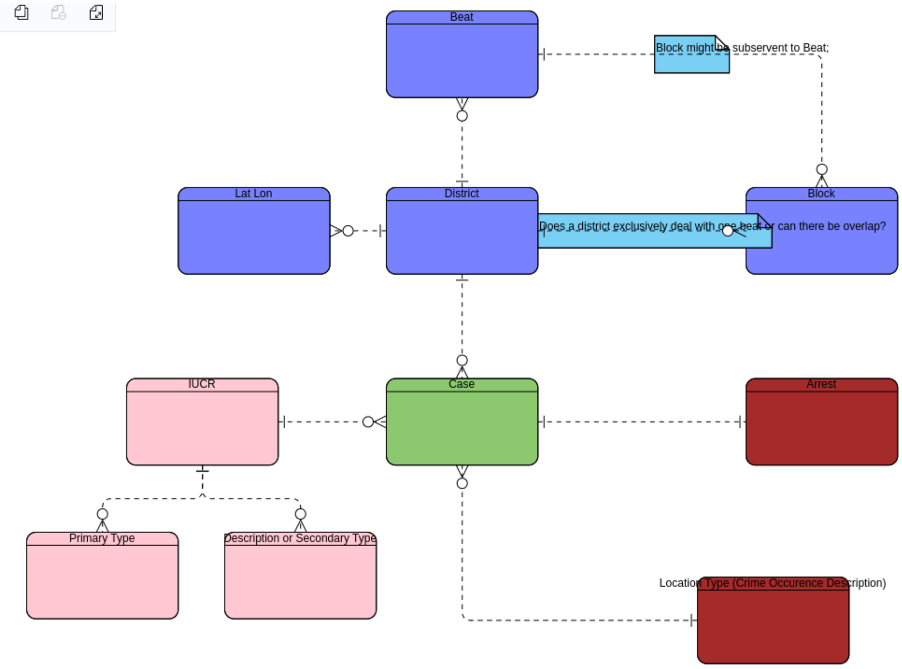
Questions: 1. What do you mean with variables? How can I even design a DB according to the questions in the first place?; eg one question is: do crime pattersn persits over time –¿ one does not design a database like this; so should I just design the database and then answer the questions with views? in my

[3] it would also be very awkward; I have never seen this case except for panel data. But this is no panel data.

opinion, one should only build Variables in Views or in Other languages. SQl is a querry language, so this does not make sense to set a variable in the DB. I rather use Views etc.

Variables Possible variables created should be used in views. Examples include aggregated crimes by district. Eg average murder per district and then further drilled down to beat; so average murder rate by beat. Then we visualize this fact. Following this we can then see beats that are really fucked up. But I would not create a varaible in the DB. **In the end we are creating a DB and not a DataWarehouse with prepared variables; Variables should be included in Views and not the database as I do not want to impose a certain variable onto a user of this DB; but views are better suited to this purpose. Databases are meant for raw data and not any transformed data**

3 Part 3- creation of an entitiy relationship diagram


Beat
Block might be subservent to Beat;
Lat Lon
District
Does a district exclusively deal with one beat or can there be overlap?
Block
IUCR
Case
Arrest
Primary Type
Description or Secondary Type
Location Type (Crime Occurence Description)