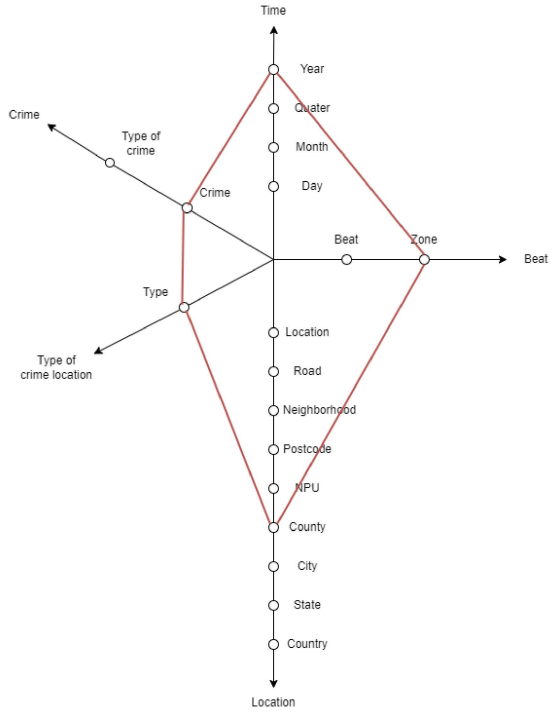
CITS3401 - Data Warehousing
Project 1 Report - Crime Investigation

Henry Tran - 23035141

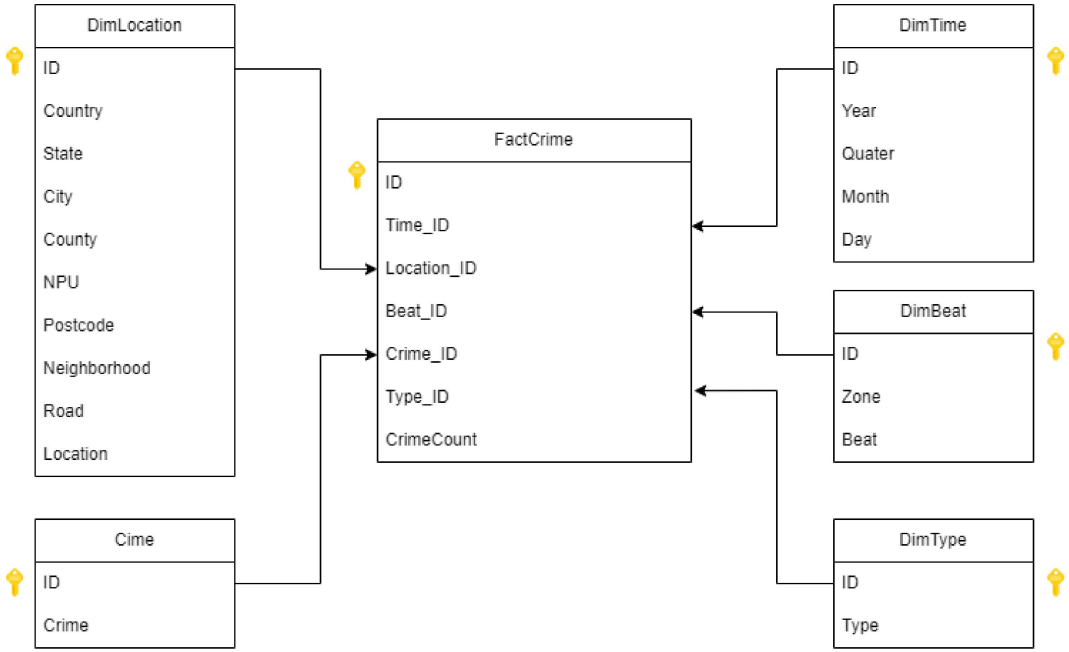
Date complete: 23/04/2023

StarNet, Fact table and dimension tables design

StarNet design



Fact table and dimension tables design



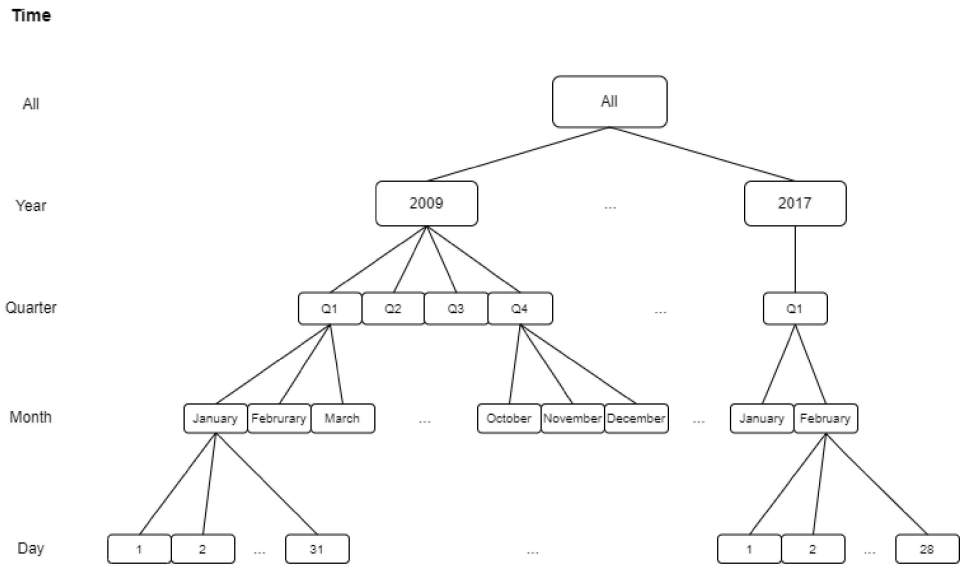
- As I have the dataset and some business questions in my mind, I have drawn the StarNet and the red footpath will illustrate how the table design can answer the questions
- The dimension tables are created based on the Starnet design and the dataset's structure.
- Some new columns in the dimension tables are calculated or extracted from the existing data, which helps answer business questions from broader perspective than the actual dataset (Eg: Zone is created from Beat, Quaters from Months)
- However, some columns which do not provide us useful information for our business questions, we will remove them.
- Each dimensions table have an unique keys and explaination of each value.
- The fact table stores unique crime events with information which are the foreign keys from the dimension tables
- Each line in the fact table represent the count of crime recorded in a specific time, location, type of crime location, beat, and crime

Schema and Concept Hierarchy

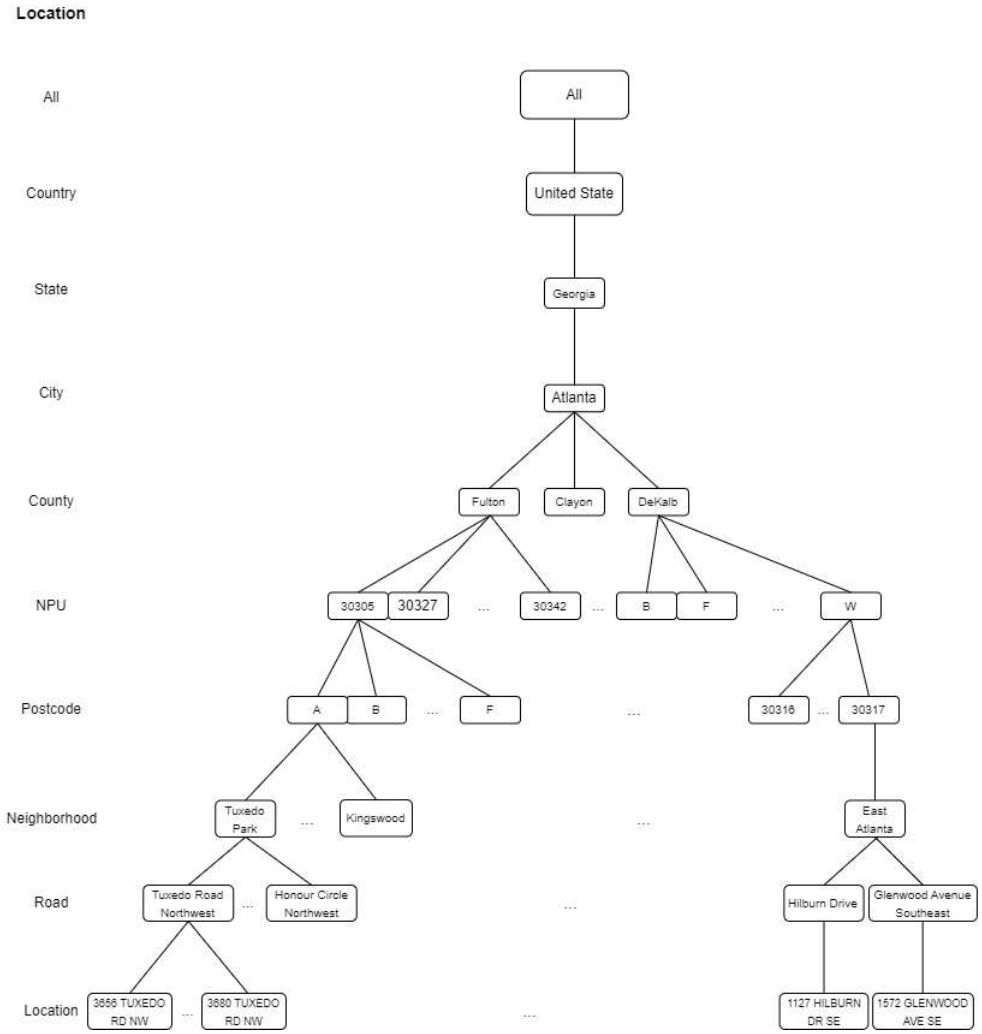
a)Time dimension and Location dimension schema and concept hierarchy designs

- The columns on the left are schema hierarchy and the diagrams on the right are concept hierarchy
- The time dimension originally got Year, Month and Day. The quarter values are generated based on the months
- The location dimension has many levels. The hierarchy can be easily made from country to county and from road to specific location. NPU, postcode and neighborhood requires more time to place them in the hierarchy.
- On the website of City of Atlanta, the city council post full information and map about city planning. It can be seen that one postcode can have one to many NPUs and each NPU will have different neighborhood. Therefore, we rank them postcode, NPU and neighborhood as from larger to smaller area.

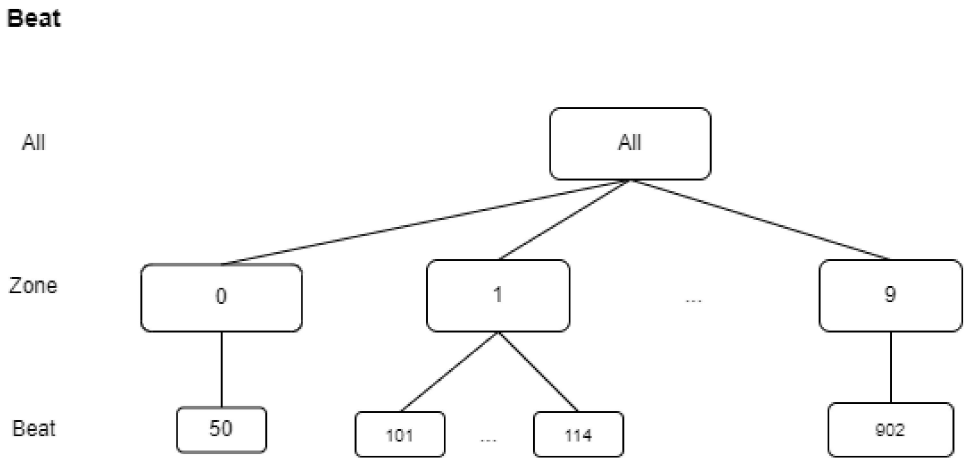
Time dimension



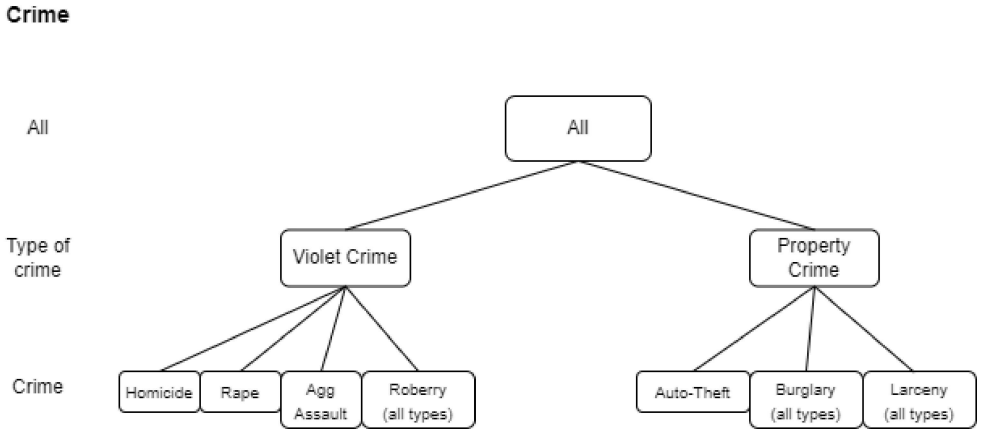
Location dimension



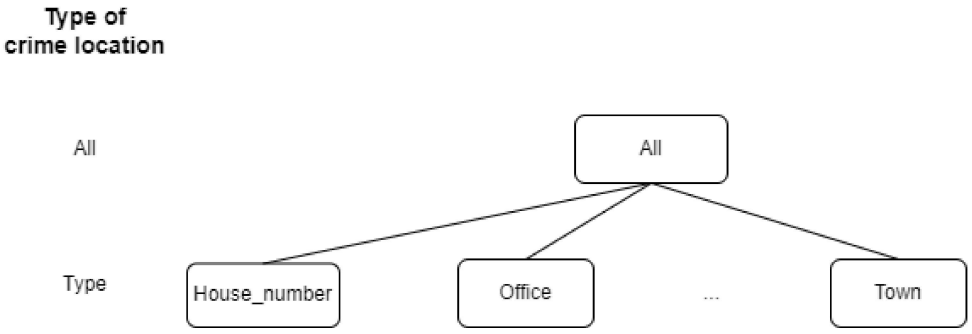
Beat dimension



Crime dimension



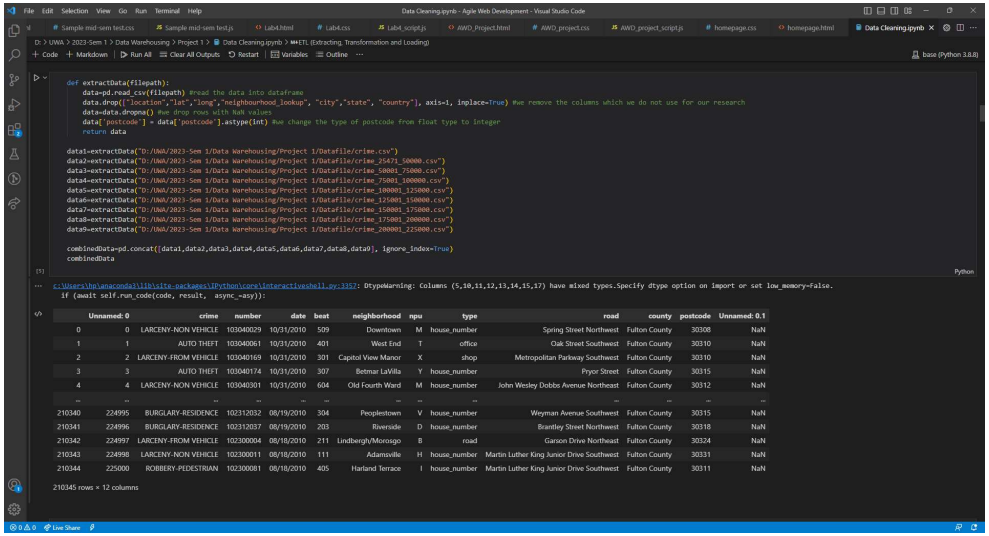
Type dimension



Schema and Concept Hierarchy

b)Beat, Crime, Type dimension schema and concept hierarchy designs

- For beat dimension, from each beat, we can know which zone it belongs to. For example, beat 101 will belong to zone 1. In our dataset, we have 9 zones.
- Looking at crime dimension, we sort the given crimes into 2 groups: violent crimes (the victims may lose their lives) and property crimes (the victims lose their property and their lives are not threatened). We sort crimes into these 2 groups based on the book " Social Problems: Continuity and Change" published by University of Minnesota Libraries Publishing (2010)
- For type dimension. we call it as "type of crime location". The reason for this change is because each value here record the type of location that a crime happened. For example, if a crime happened in a specific location and that location is an office, then the type of crime location is office



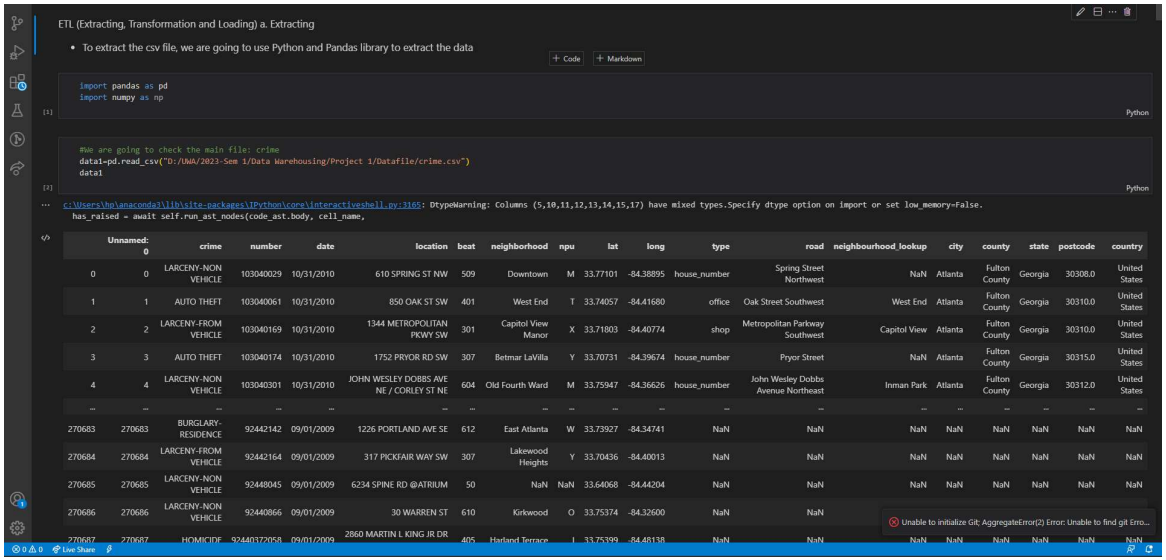
ETL process

- We will use Python for the ETL process.
- In extraction the data, we look through the data in the file "crime" by checking the value in each column, remove invalid values and merge it with other datafile (we will use the whole dataset for this project)
- After the extraction process, we transform the data into dimension and fact tables based on our dimensions that we drawn in our Starnet diagram.
- Then, we extract each dimension table and fact table into ".csv" file and we load it into SSMS to create database and manage it.

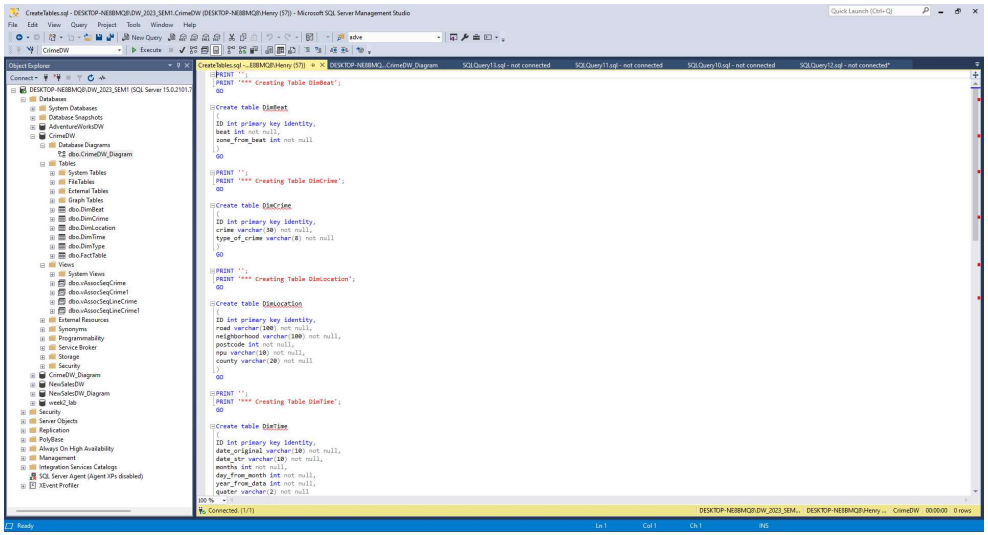
- When we open SSMS, we create database, create table, create foreign keys and load the data into our database (each task we have its own SQL script for it)
- We create a database diagram to visualize the connection between fact table with other dimension tables.
- Another method we can do is we create the diagram first and then we load the data in after that by writing SQL script (we can use the same SQL upload script for both method)
- After finishing with the diagram, we can perform roll-up/ drill-down to see the result. We can also see roll-up/drill-down result in multidimensional data cubes in SSDT.

- We open the Visual Code Studio 2019 to make the data cubes by selecting "Analysis Service Multidimensional and Data Mining Project".
- Then, we create the data source view, data cubes and generate the hierarchy for each dimension.
- In here, we can check the roll-up/ drill-down analysis of the data cube

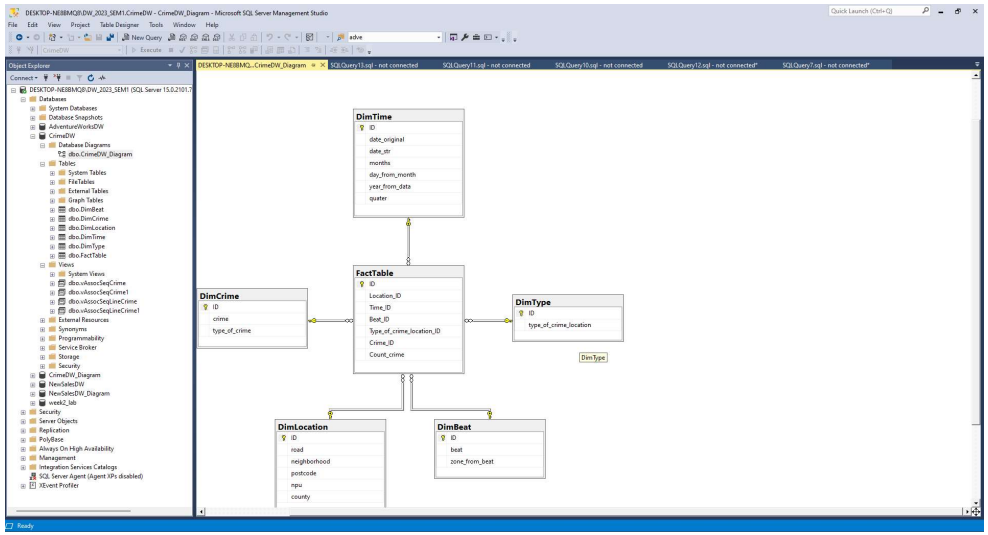
- The last step we will do it to visualize our data using PowerBI to answer our business queries



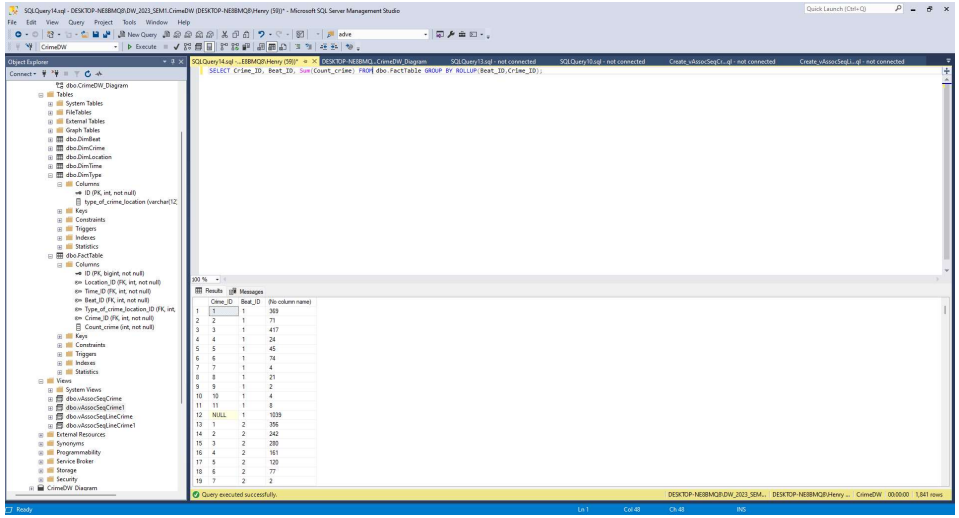
Create Table Script



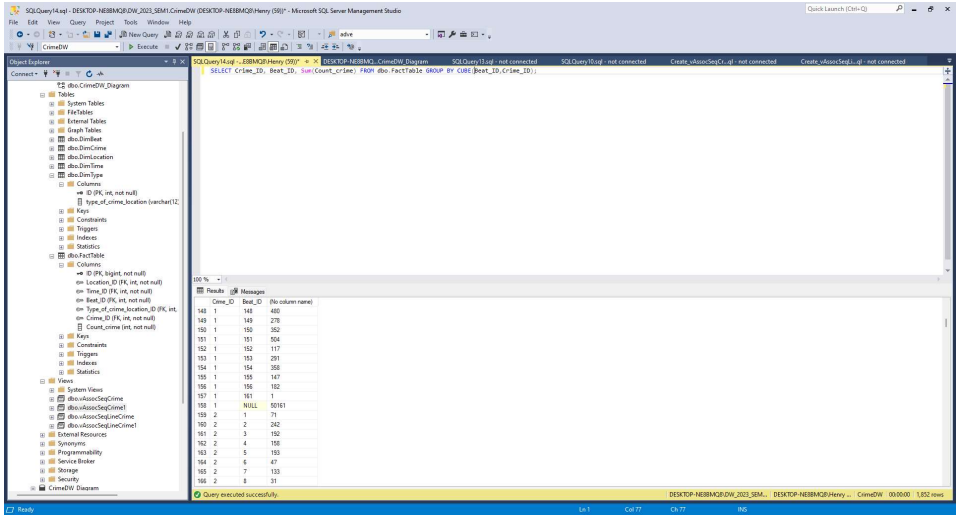
Schema Diagram on SSMS



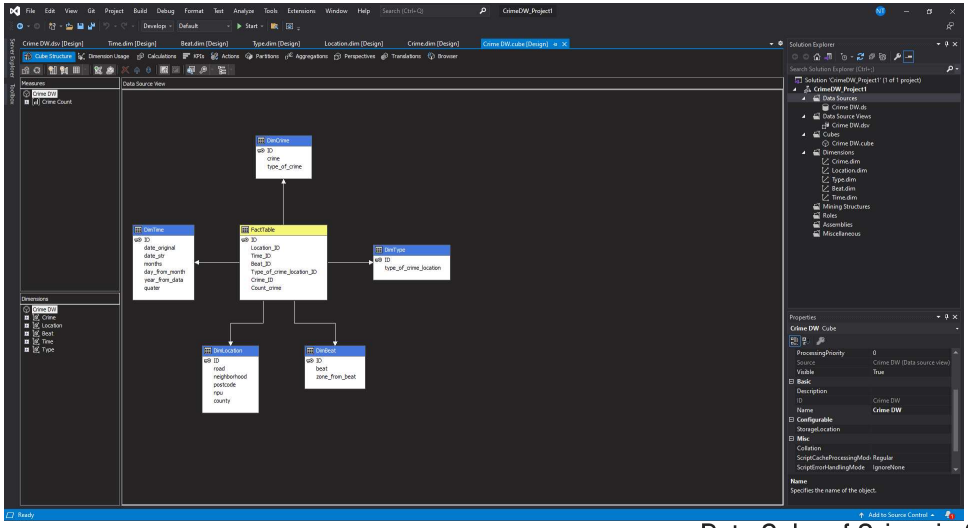
Roll-up with Crime_ID and Beat_ID



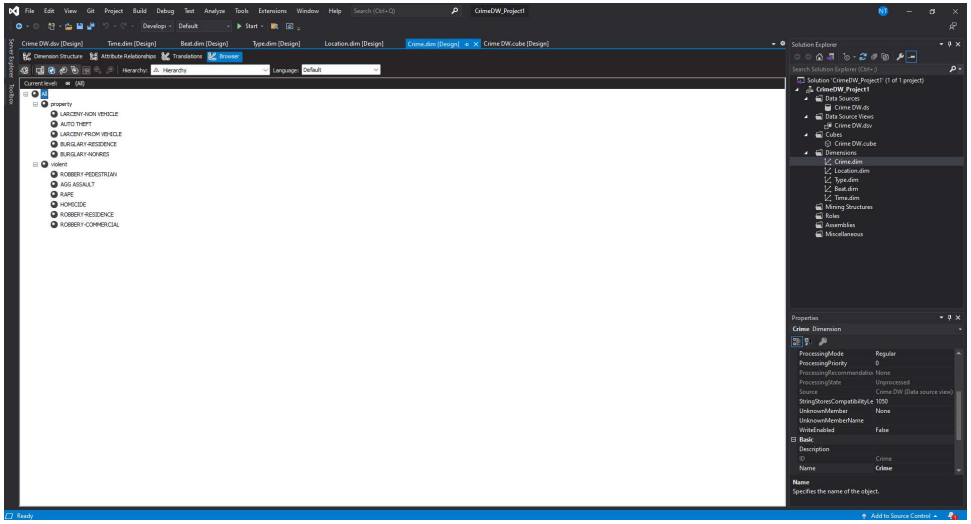
Cube using Crime_ID and Beat_ID



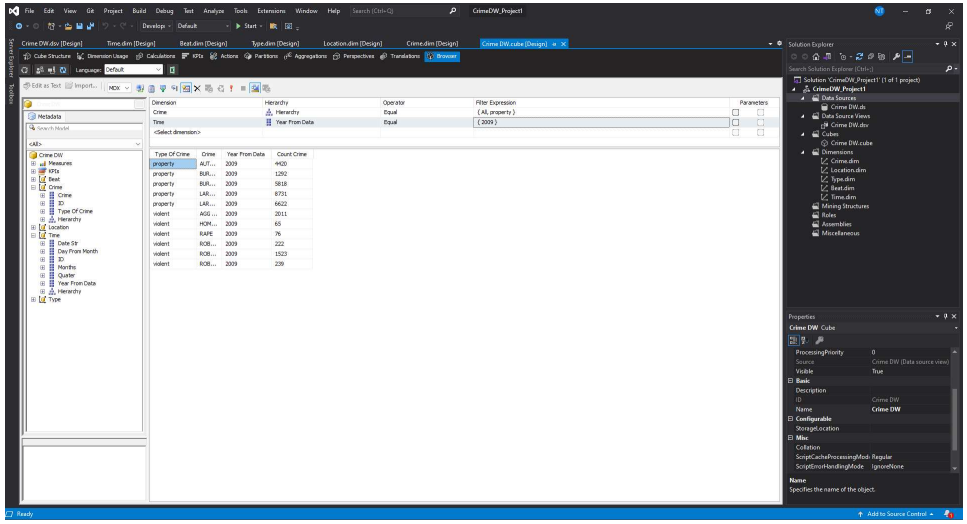
Data Cube Diagram



Crime Hierarchy



Data Cube of Crime in 2009



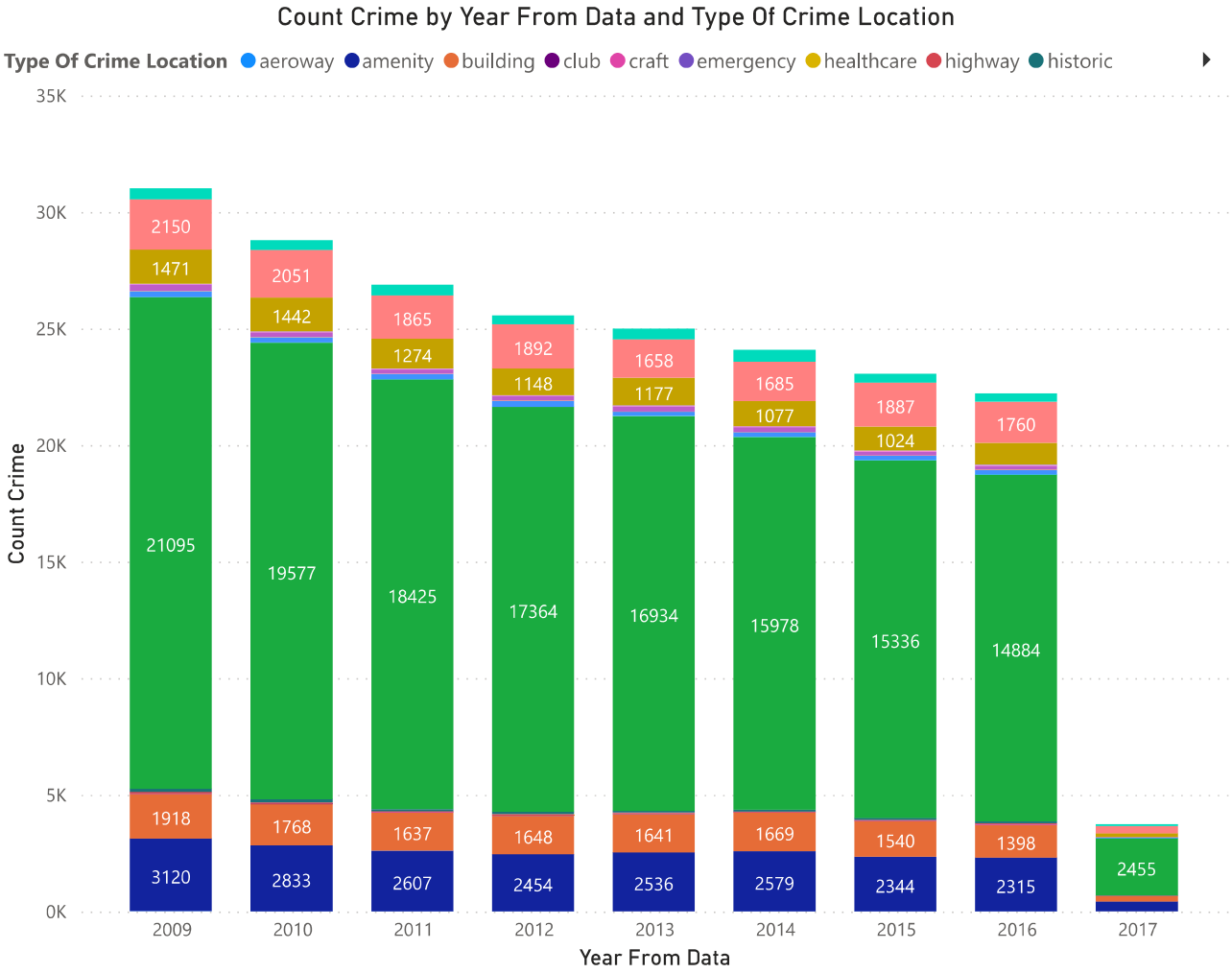
Business queries

Q1: What is the change in the number of crime through each year?

- Overall we can see the number of crime decrease through the year, from more than 30,000 crime records in 2009 to around 22,500 records in 2016.
- Year 20017, we can see that the number of crime records is significantly lower than many previous year. The reason for this change is because there are only the crime records of January and February in this year.

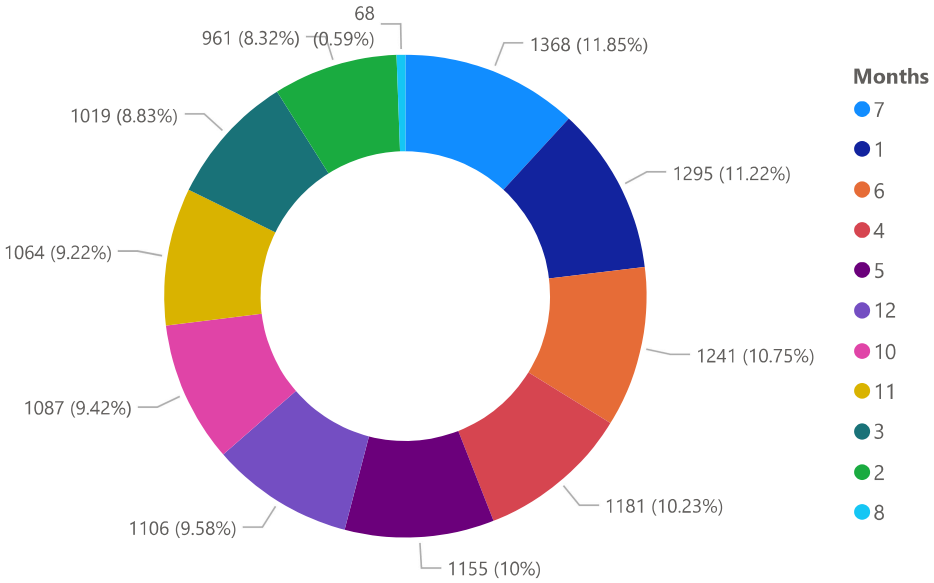
Q2: Which type of crime locations occurs in large numbers each year?

- There are 19 type of location that the crimes take place. All the crime locations decreased the number of records through 8 years.
- However, we have 5 location types that the crimes take place most in all the year in the data set: house_number, amenity, building, shop, road.
- For house_number gets the highest number of records in every years, even just within 2 months of 2017.

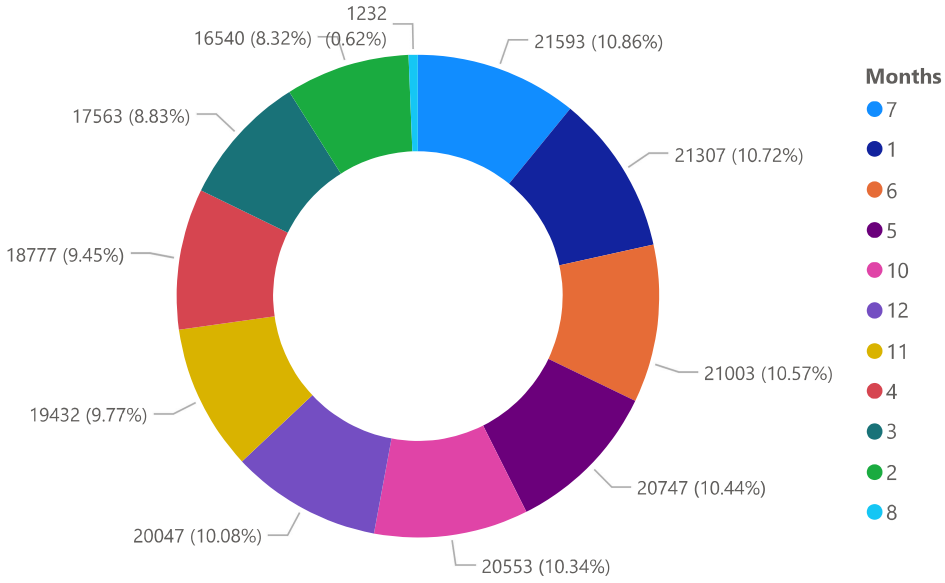


Business queries

Number of crime in Dekalb county in each months

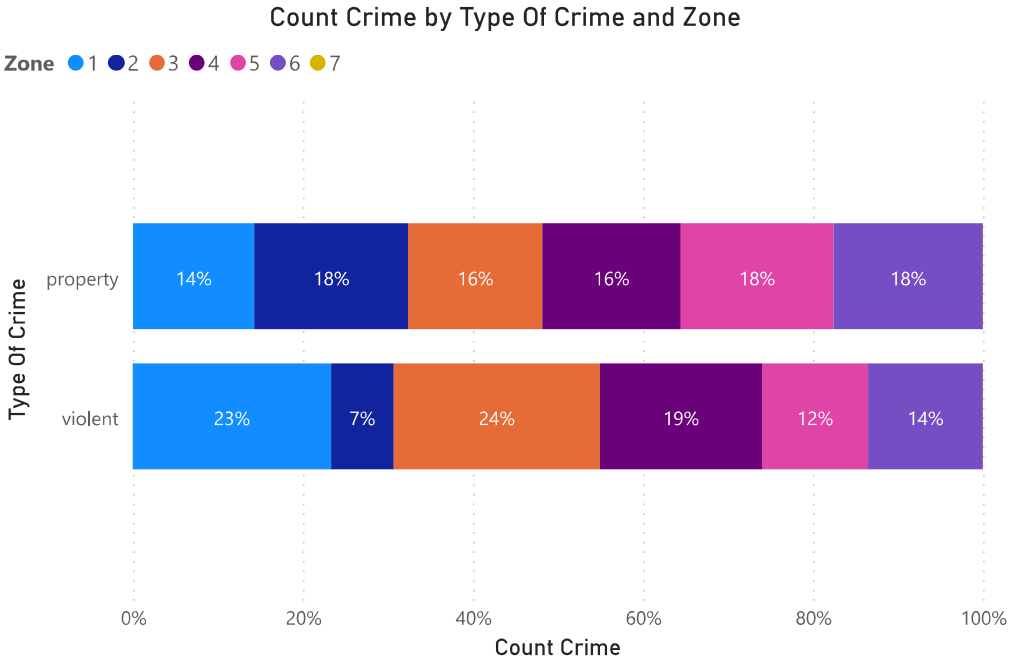


Number of crime in Fulton county in each months



Q3: What is the number of crime in different counties in different months of all the years?

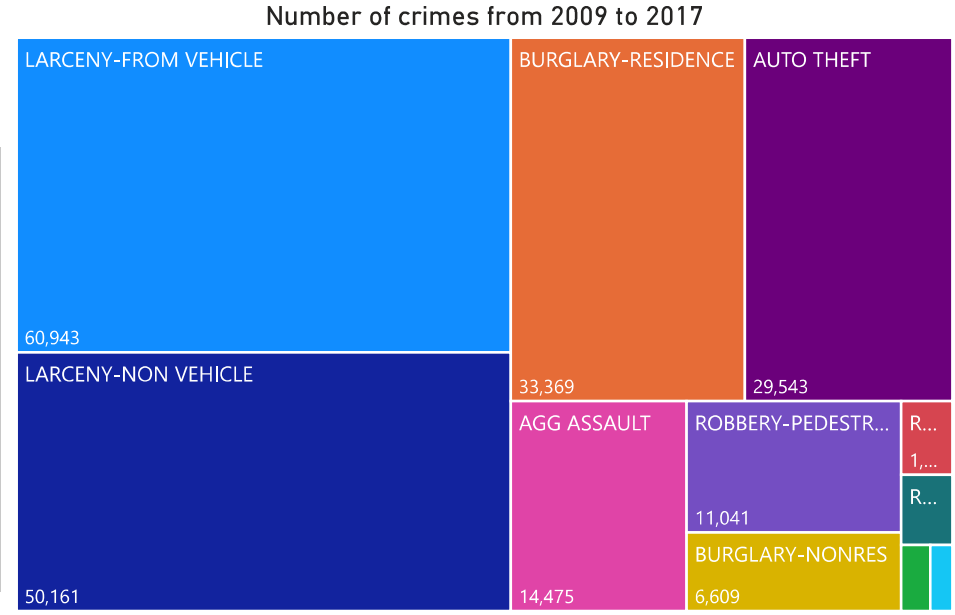
- We are having 2 graphs which represent the number of crime in different months from 2009 to 2017 in Dekalb and in Fulton.
- Firstly, there are only 11 months in the graph due to the missing of crime records in September. There are some reasons that can lead to the missing of September in our analysis, for example: the records in September did not have enough information to be counted as valid records, etc.
- Secondly, even though having 2 different areas, 2 different number of crimes, we can see the same pattern in the number of crime records. The number of records is high in some months in the middle of the year, January and follows by the last 3 months of the year. February, March and August are 3 months having the percentages of crime record lower than 9%.



Q4: Which zone has the the highest rates of crimes in each type of crime ?

- The graph on the left shows the percentage of crimes in different zones based on crime types
- There are 2 different crimes: property crimes and violent crime type.
- With property crime, the percentage of crime cases are almost similar to each other, around 16% to 18%, except zone number one with 14%.
- For violent crime, the zone has the highest percentages of crime records in this type is zone 3 with 24%, follow by zone 1 and zone 4 ranked the second and third with 23% and 19% respectively. The other 3 zones have lower percentage of violent crimes recorded.
- Zone 7 takes really small percentage of crimes in both crime types

(The drill down can be perform to see the percentage of each crime recorded in each zone which will support the question 5)



Q5: Which crimes have the highest number of records and which zone has the highest/lowest rate of that crime within 2009-2017?

- The graph used to illustrate the data is treemap, we can see the magnitude of each crimes.
- Overall, we can see that the number of property crime recorded is higher than the number of violent crimes.
- Larceny in all kinds has the largest number of cases, with almost 61 thousand cases for larceny from vehicle and 51.1 thousand cases for larceny non vehicle.
- Homicide and rape have the smallest number of cases.

(After clicking drill drown from the graph in question 4, clicking in the crime in graph of question 5 to see the percentage of cases in different zone with the selected crime)

Reference

City of Atlanta. (n.d.). *Atlanta, GA*. City of Atlanta, GA. https://www.atlantaga.gov/government/departments/city-planning/maps-and-gis

University of Minnesota Libraries Publishing. (2016, March 25). *8.2 types of crime – Social problems*. https://open.lib.umn.edu/socialproblems/chapter/8-2-types-of-crime/#