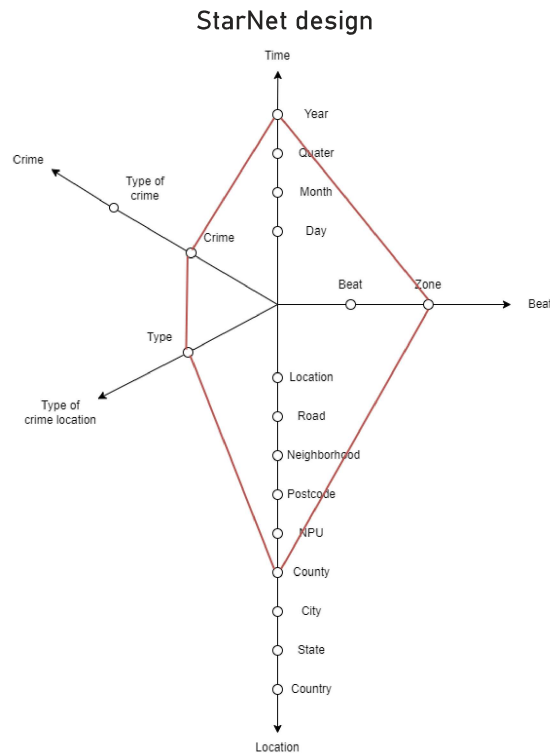


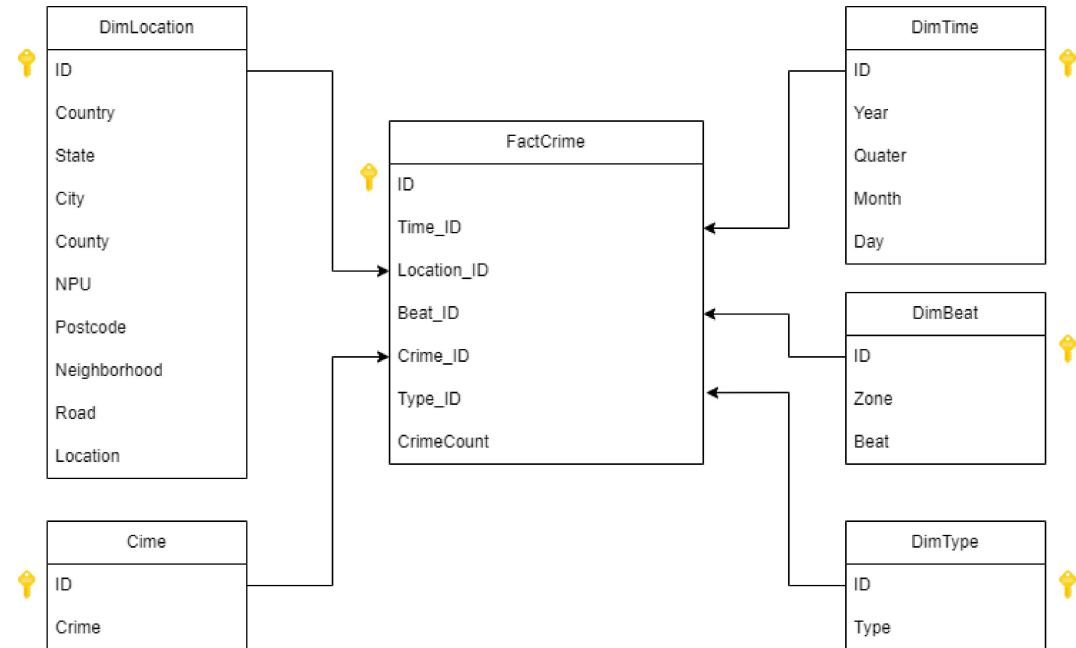
CITS3401 - Data Warehousing
Project 1 Report - Crime Investigation

Henry Tran - 23035141
Date complete: 23/04/2023

StarNet, Fact table and dimension tables 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 footprint 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 explanation 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

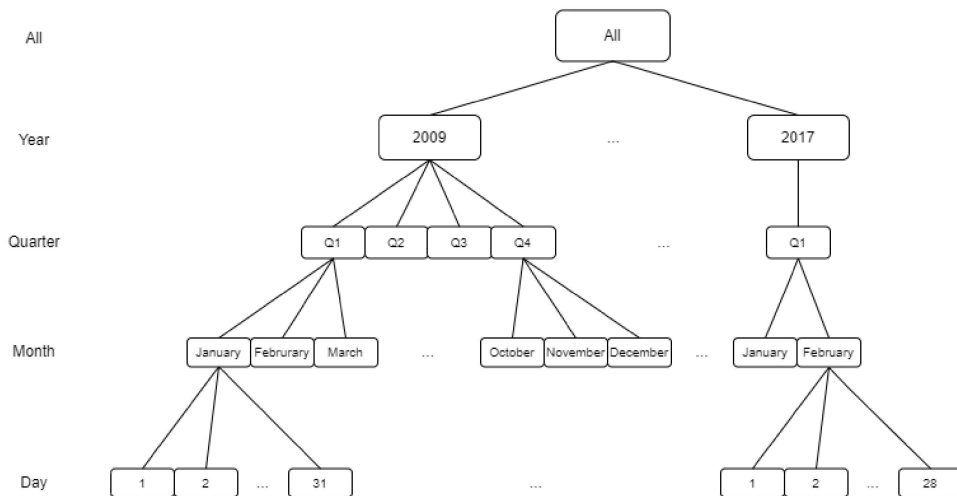
Schema and Concept Hierarchy

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.

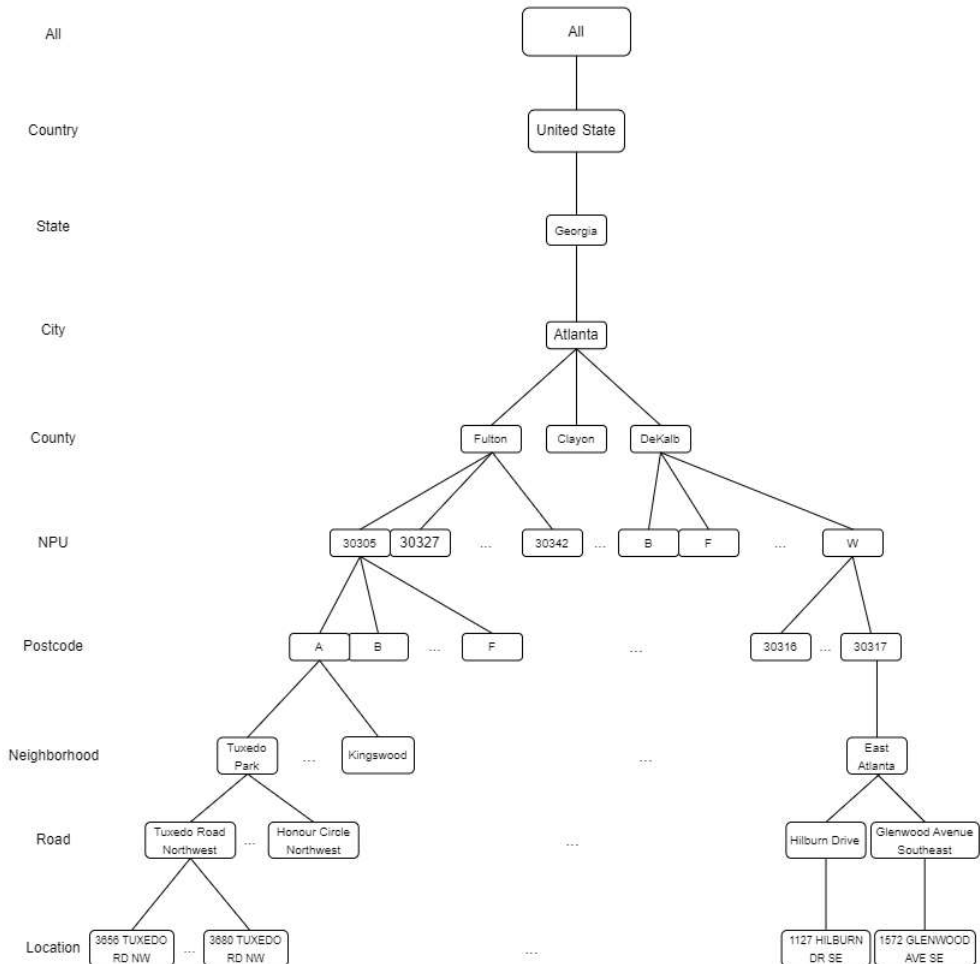
Time dimension

Time



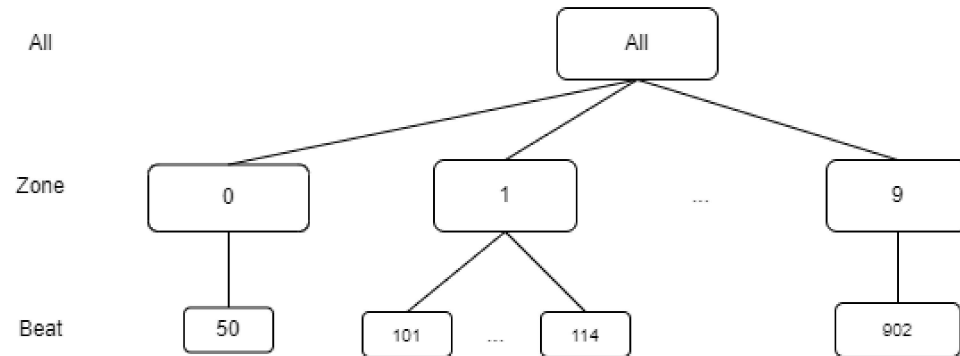
Location dimension

Location



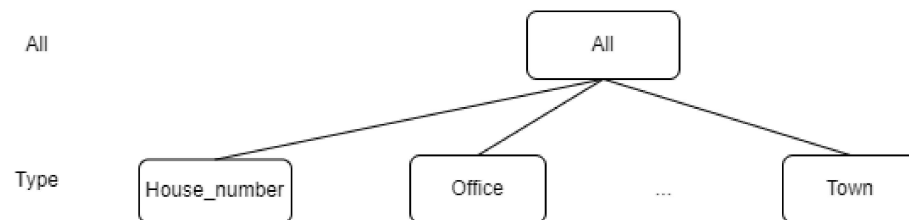
Beat dimension

Beat



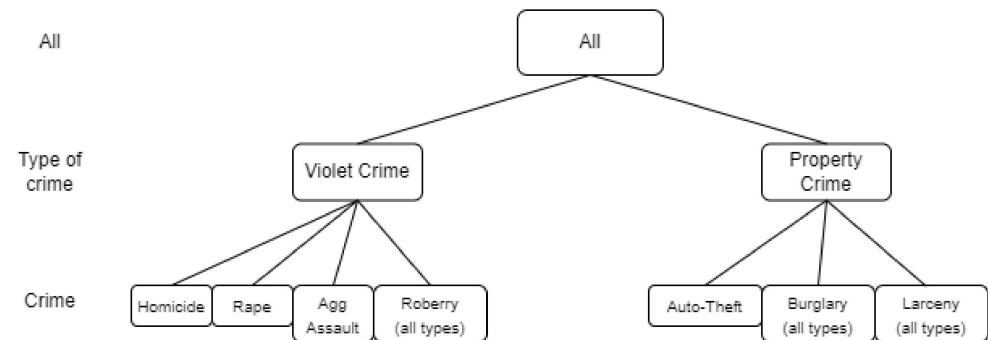
Type dimension

Type of
crime location



Crime dimension

Crime



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

```
def extractData(filepath):
    data=pd.read_csv(filepath)
    data.drop(["location","lat","long","neighbourhood_lookup","city","state"],axis=1,inplace=True) #we remove the columns which we do not use for our research
    data=data.dropna() #we drop rows with NaN values
    data["postcode"] = data["postcode"].astype(int) #we change the type of postcode from float type to integer
    return data

data1=extractData("D:/Data Warehouse/Project 1/Datafile/crime.csv")
data2=extractData("D:/Data Warehouse/Project 1/Datafile/crime_25471_50000.csv")
data3=extractData("D:/Data Warehouse/Project 1/Datafile/crime_50001_75000.csv")
data4=extractData("D:/Data Warehouse/Project 1/Datafile/crime_75001_100000.csv")
data5=extractData("D:/Data Warehouse/Project 1/Datafile/crime_100001_125000.csv")
data6=extractData("D:/Data Warehouse/Project 1/Datafile/crime_125001_150000.csv")
data7=extractData("D:/Data Warehouse/Project 1/Datafile/crime_150001_175000.csv")
data8=extractData("D:/Data Warehouse/Project 1/Datafile/crime_175001_200000.csv")
data9=extractData("D:/Data Warehouse/Project 1/Datafile/crime_200001_225000.csv")

combinedData=pd.concat([data1,data2,data3,data4,data5,data6,data7,data8,data9], ignore_index=True)
combinedData
```

Unnamed: 0	crime	number	date	beat	neighbourhood	npu	type	road	county	postcode	Unnamed: 0.1	
0	LARCENY-NON VEHICLE	103040029	10/31/2010	509	Downtown	M	house_number	Spring Street Northwest	Fulton County	30308	NaN	
1	AUTO THEFT	103040061	10/31/2010	401	West End	T	office	Oak Street Southwest	Fulton County	30310	NaN	
2	LARCENY-FROM VEHICLE	103040169	10/31/2010	301	Capitol View Manor	X	shop	Metropolitan Parkway Southwest	Fulton County	30310	NaN	
3	AUTO THEFT	103040174	10/31/2010	307	Betmar LaVilla	Y	house_number	Pryor Street	Fulton County	30315	NaN	
4	LARCENY-NON VEHICLE	103040301	10/31/2010	604	Old Fourth Ward	M	house_number	John Wesley Dobbs Avenue Northeast	Fulton County	30312	NaN	
210340	224995	BURGLARY-RESIDENCE	103212052	09/19/2010	304	Peoplesstown	V	house_number	Weyman Avenue Southwest	Fulton County	30315	NaN
210341	224996	BURGLARY-RESIDENCE	103212057	09/19/2010	203	Riverside	D	house_number	Bradley Street Northwest	Fulton County	30318	NaN
210342	224997	LARCENY-FROM VEHICLE	103200004	09/16/2010	211	Lindbergh'sMemorial	B	road	Gordon Drive Northeast	Fulton County	30324	NaN
210343	224998	LARCENY-NON VEHICLE	103200011	09/16/2010	111	Adamsville	H	house_number	Martin Luther King Junior Drive Southwest	Fulton County	30321	NaN
210344	225000	ROBBERY-FLEXISTRAIN	103200001	09/16/2010	405	Harland Terrace	I	house_number	Martin Luther King Junior Drive Southwest	Fulton County	30311	NaN

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 Starret 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.

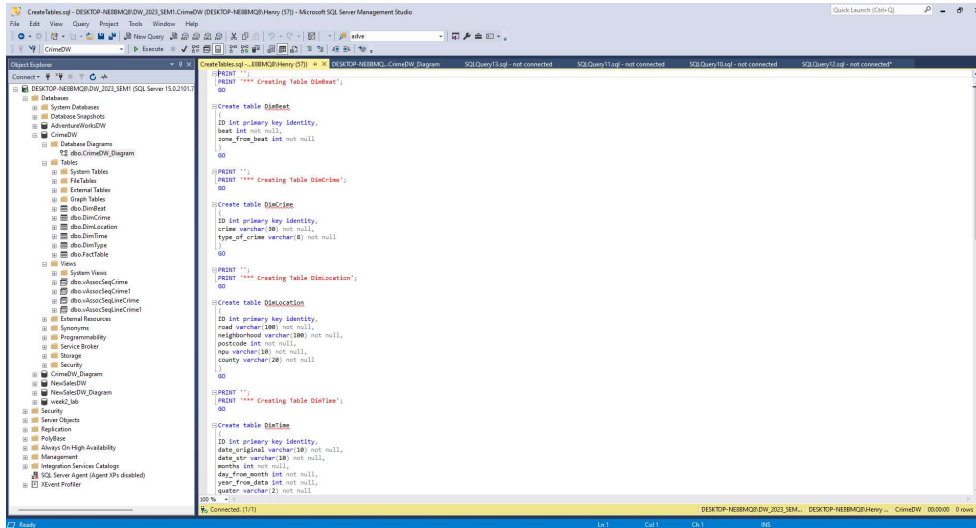
- We open the Visual Code Studio 2019 to make the data cubes by selecting "Analysis Service Multidimensional and Data Mining Project".
- Then, we create the data source view, data cubes and generate the hierarchy for each dimension.
- In here, we can check the roll-up/ drill-down analysis of the data cube
- The last step we will do it to visualize our data using PowerBI to answer our business queries

```
import pandas as pd
import numpy as np

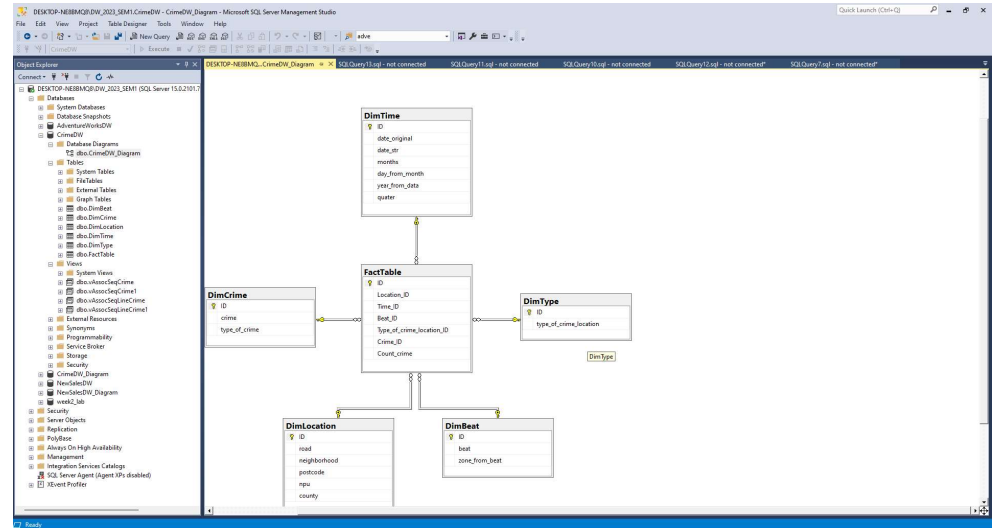
#we are going to check the main file: crime
data1=pd.read_csv("D:/Data Warehouse/Project 1/Datafile/crime.csv")
data1
```

Unnamed: 0	crime	number	date	location	beat	neighbourhood	npu	lat	long	type	road	neighbourhood_lookup	city	county	state	postcode	country
0	LARCENY-NON VEHICLE	103040029	10/31/2010	610 SPRING ST NW	509	Downtown	M	33.77101	-84.38895	house_number	Spring Street Northwest	NaN	Atlanta	Fulton County	Georgia	30308.0	United States
1	AUTO THEFT	103040061	10/31/2010	850 OAK ST SW	401	West End	T	33.74057	-84.41680	office	Oak Street Southwest	West End	Atlanta	Fulton County	Georgia	30310.0	United States
2	LARCENY-FROM VEHICLE	103040169	10/31/2010	1344 METROPOLITAN PKWY SW	301	Capitol View Manor	X	33.71803	-84.40774	shop	Metropolitan Parkway Southwest	Capitol View	Atlanta	Fulton County	Georgia	30310.0	United States
3	AUTO THEFT	103040174	10/31/2010	1752 PRYOR RD SW	307	Betmar LaVilla	Y	33.70731	-84.39674	house_number	Pryor Street	NaN	Atlanta	Fulton County	Georgia	30315.0	United States
4	LARCENY-NON VEHICLE	103040301	10/31/2010	JOHN WESLEY DOBBS AVE NE / CORLEY ST NE	604	Old Fourth Ward	M	33.75947	-84.36626	house_number	John Wesley Dobbs Avenue Northeast	Inman Park	Atlanta	Fulton County	Georgia	30312.0	United States
270683	270683	BURGLARY-RESIDENCE	92442142	09/01/2009	1226 PORTLAND AVE SE	612	East Atlanta	W	33.73927	-84.34741	NaN	NaN	NaN	NaN	NaN	NaN	NaN
270684	270684	LARCENY-FROM VEHICLE	92442164	09/01/2009	317 PICKFAIR WAY SW	307	Lakewood Heights	Y	33.70436	-84.40013	NaN	NaN	NaN	NaN	NaN	NaN	NaN
270685	270685	LARCENY-NON VEHICLE	92448045	09/01/2009	6234 SPINE RD @ATLANTA	50	NaN	NaN	33.64068	-84.44204	NaN	NaN	NaN	NaN	NaN	NaN	NaN
270686	270686	LARCENY-NON VEHICLE	92448066	09/01/2009	30 WARREN ST	610	Kirkwood	O	33.75374	-84.32600	NaN	NaN	NaN	NaN	NaN	NaN	NaN
270687	270687	HOMICIDE	92448172058	09/01/2009	2860 MARTIN L KING JR DR	405	Harland Terrace	I	33.75393	-84.41138	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Create Table Script



Schema Diagram on SSMS



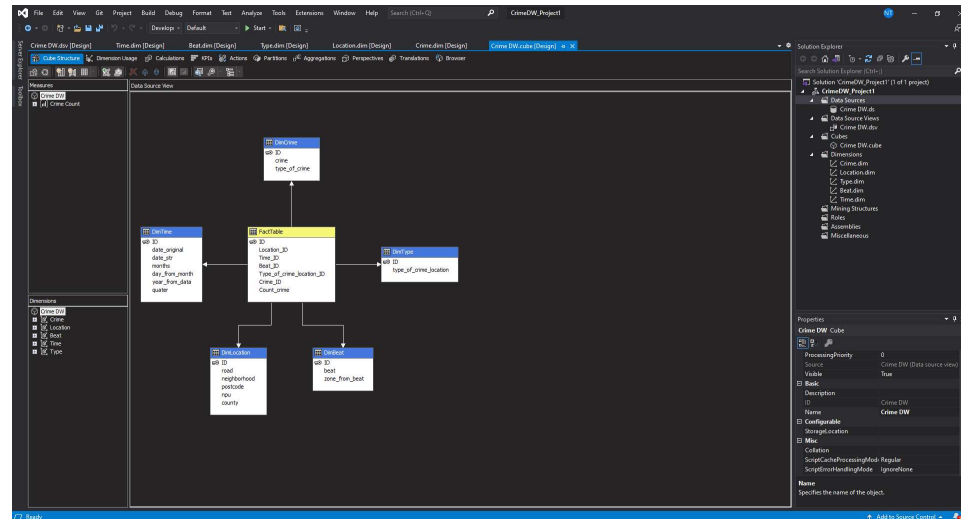
Roll-up with Crime_ID and Beat_ID

Crime_ID	Beat_ID	sum(Crime_Count)
1	1	389
2	1	71
3	1	437
4	1	24
5	1	45
6	1	74
7	1	4
8	1	23
9	1	2
10	1	4
11	1	8
12	1	1039
13	1	2
14	2	242
15	2	389
16	2	161
17	2	135
18	2	77
19	2	2

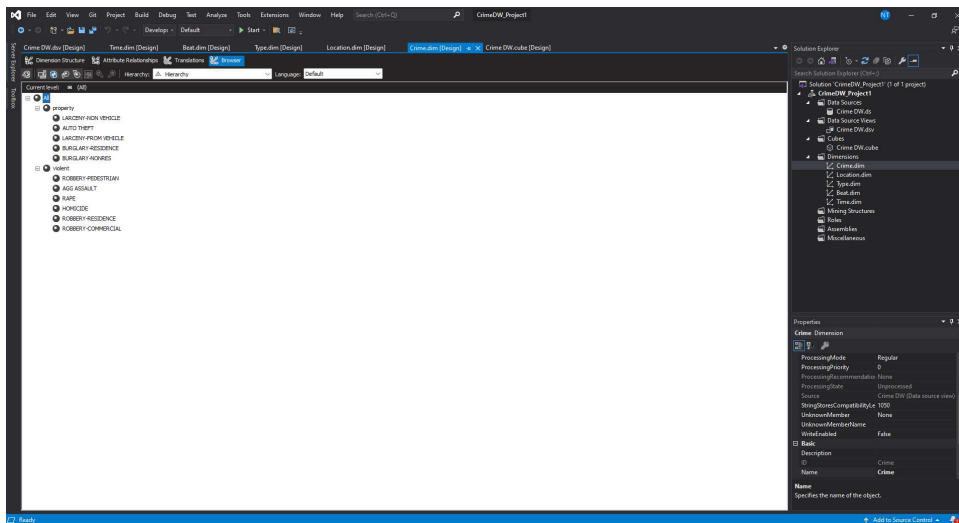
Cube using Crime_ID and Beat_ID

Crime_ID	Beat_ID	sum(Crime_Count)
1	1	460
140	1	276
150	1	262
151	1	504
152	1	117
153	1	291
154	1	358
155	1	147
156	1	192
157	1	161
158	1	5061
159	2	71
160	2	242
161	2	162
162	2	168
163	2	193
164	2	47
165	2	133
166	2	31

Data Cube Diagram



Crime Hierarchy



Data Cube of Crime in 2009

The screenshot displays the SQL Server Data Tools interface with the Data Cube of Crime in 2009. The cube is displayed in the Solution Explorer, showing the relationship between Crime, Location, and Time dimensions. The cube is structured as follows: Crime (parent) -> Location (child) -> Time (child). The cube is also displayed in the Properties window, showing the relationship between the dimensions and measures.

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Crime	Hierarchy	Equal	{All property}	
Time	Year From Date	Equal	{2009}	

Type Of Crime	Crime	Year From Date	Count Crime
property	ALL...	2009	4020
property	BAR...	2009	1202
property	BAR...	2009	9818
property	LAR...	2009	8731
property	LAR...	2009	6622
violent	AGG...	2009	2011
violent	HOM...	2009	45
violent	RAPE	2009	76
violent	ROB...	2009	221
violent	ROB...	2009	1322
violent	ROB...	2009	239

Business queries

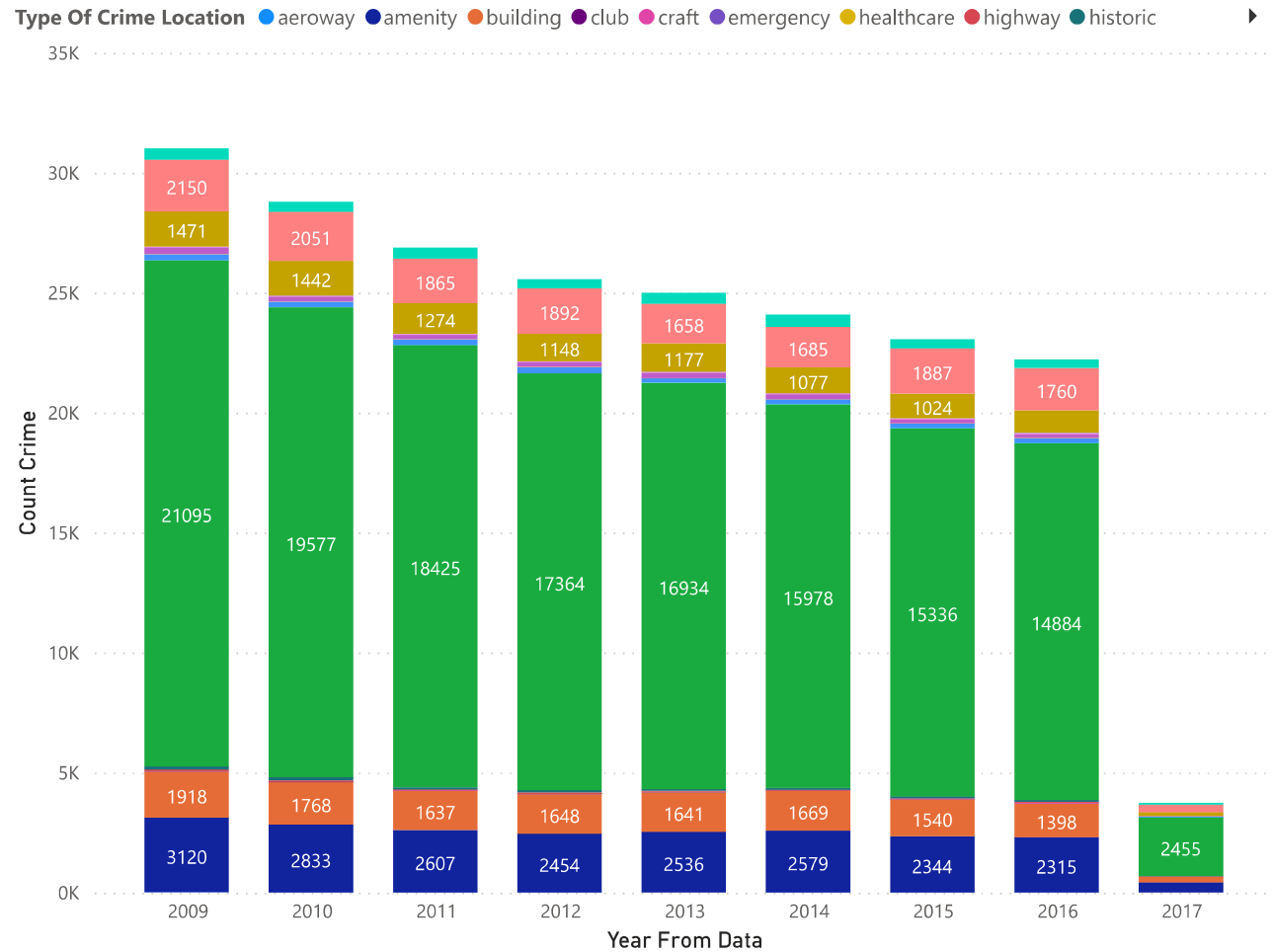
Q1: What is the change in the number of crime through each year?

- 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?

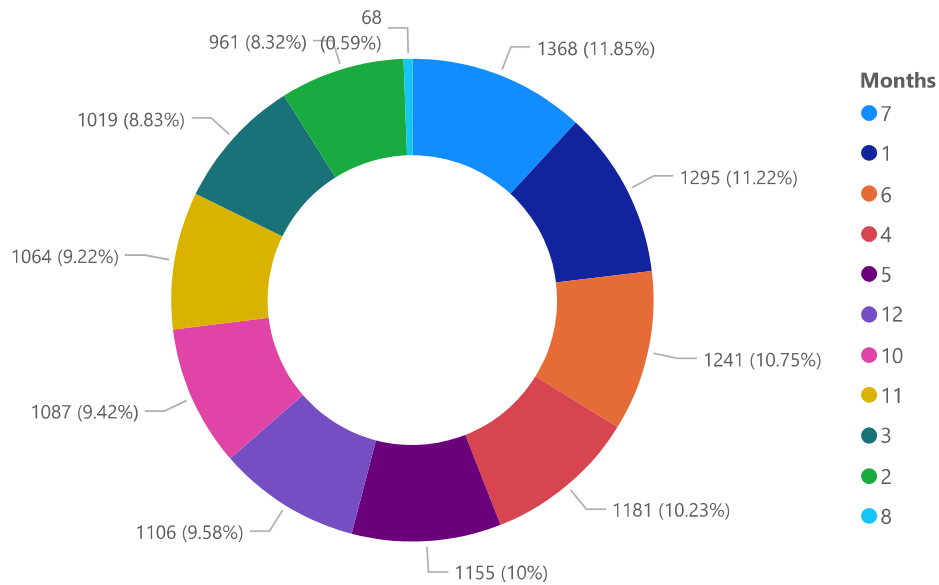
- There are 19 type of location that the crimes take place. All the crime locations decreased the number of records through 8 years.
- However, we have 5 location types that the crimes take place most in all the year in the data set: house_number, amenity, building, shop, road.
- For house_number gets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

Count Crime by Year From Data and Type Of Crime Location

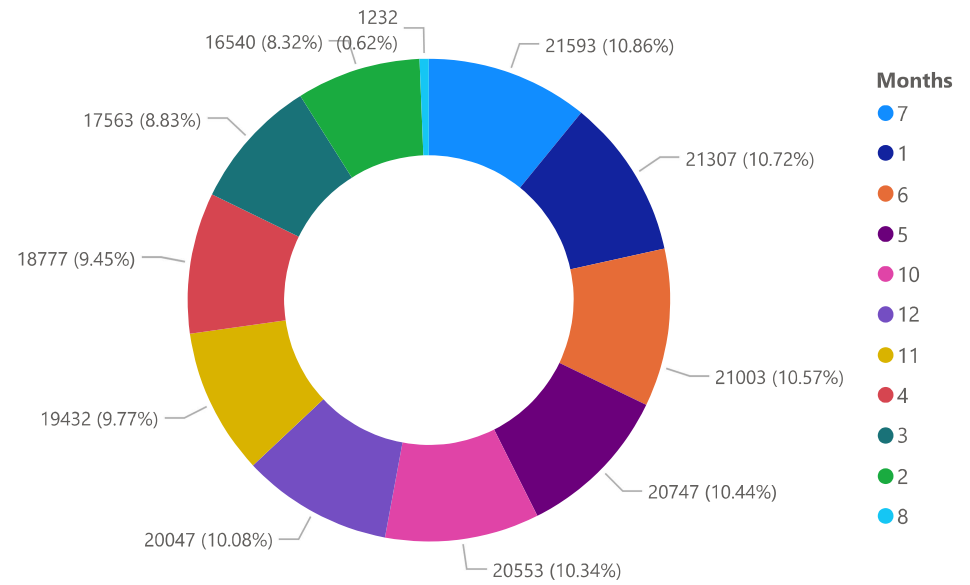


Business queries

Number of crime in Dekalb county in each months



Number of crime in Fulton county in each months

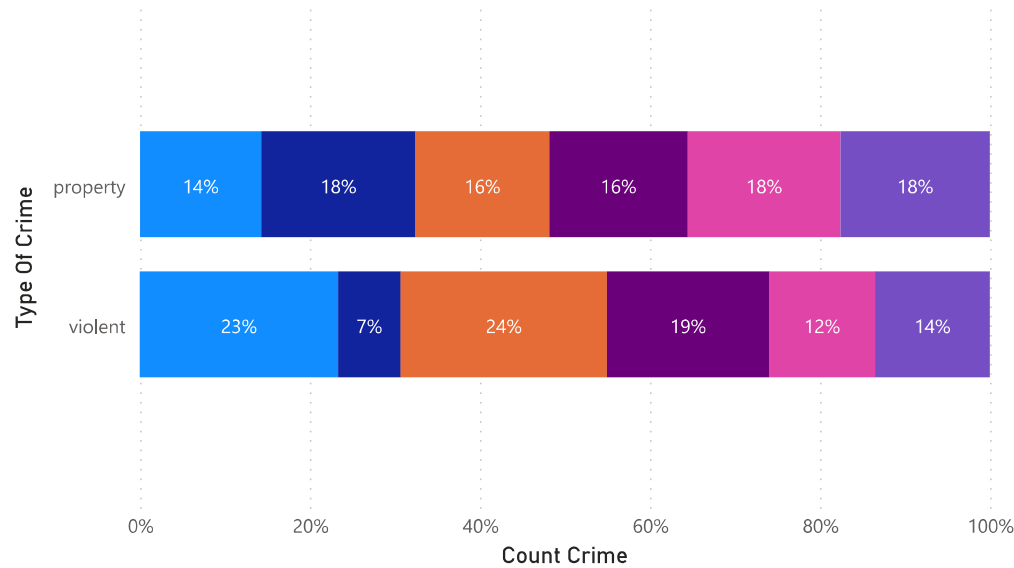


Q3: What is the number of crime in different counties in different months of all the years?

- 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%.

Count Crime by Type Of Crime and Zone

Zone 1 2 3 4 5 6 7



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

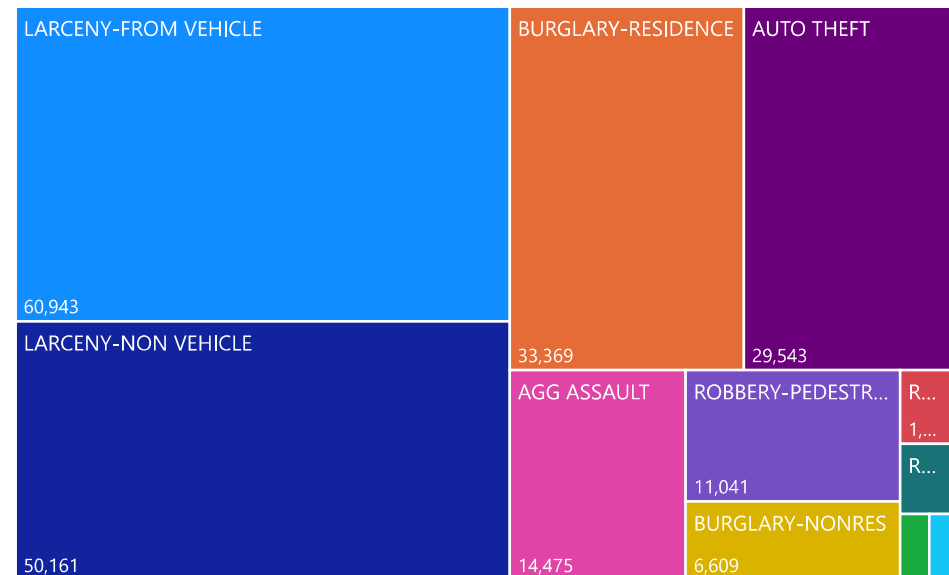
(The drill down can be perform to see the percentage of each crime recorded in each zone which will support the question 5).

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 down 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).

Number of crimes from 2009 to 2017



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/#>