Project 1 : Data Warehousing with Atlanta Crime Data

Franco Meng, 2023, April.

Abstract :

The raw dataset of 225,000 cases of crime records, is based on the Crime in Atlanta, Georgia, US. (2009 - 2017) . Retrieved from data.world: Link

The following process has been performed, to discover useful insight based on all the records of the crime occurrences over the years, also to better facilitator police department for the future planning, in order to lower the crime rate, and to keep the citizens safe.

Process/Methods/Tools :

- Design:
  Kimball's four steps, StarNet, Star Schema, ER diagram
- ETL Process:
  Python, SQL
- Server, Data model, Data cube design :
  SSMS, SSDT, SSAS, Visual Studio
- Analysis & Visualization :
  Power BI
- Data Mining, Association Rules :
  SSAS Engine

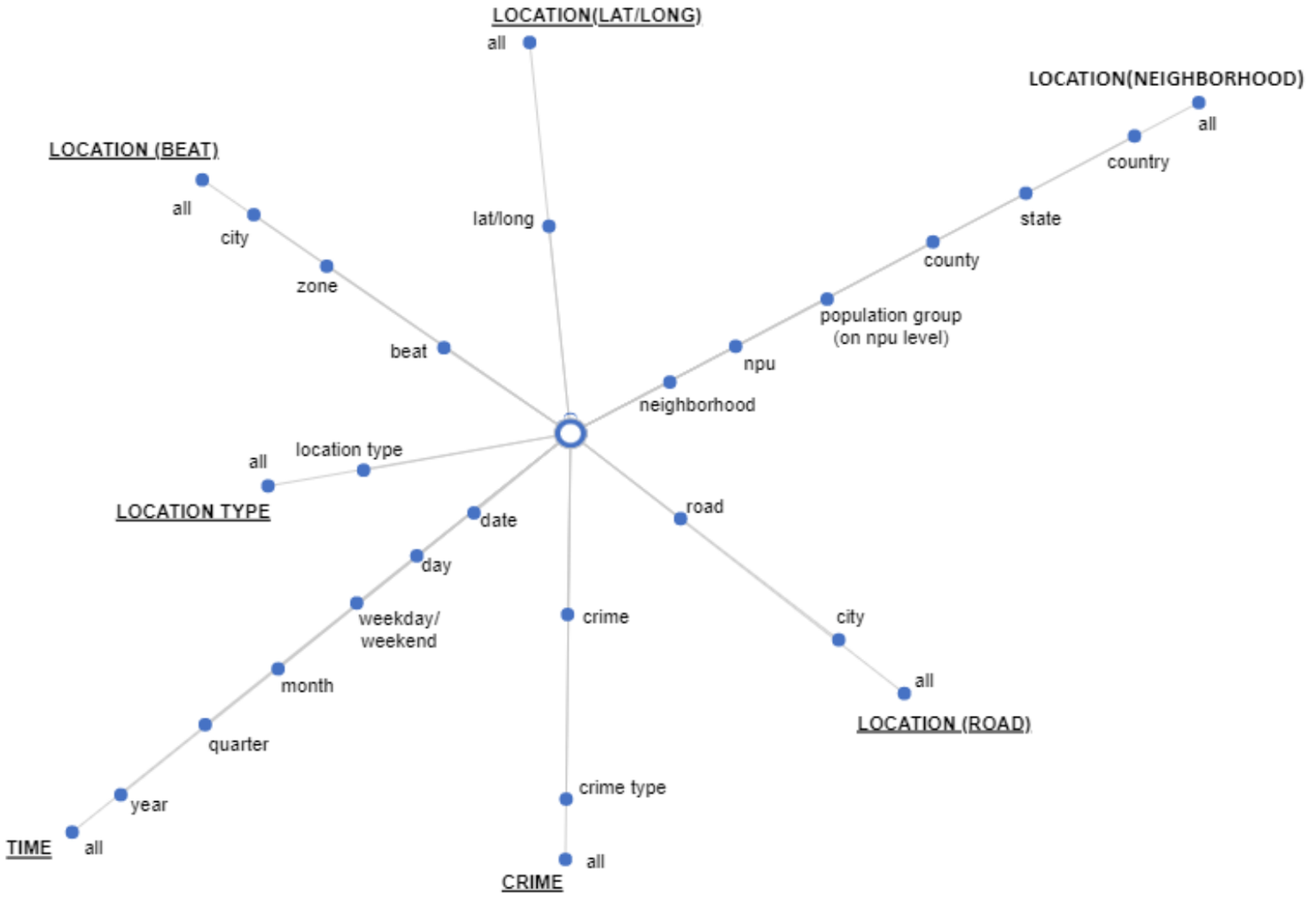
Results : *** Only a few listed here, the project is capable to answer various enquiries based on skate holder interests.**

1. One location, (33.84676, -87.36212) Lenox Square Shopping Mall . Had the highest 2946 crimes recorded, counting 1.31% of all the 225,000 records, out of 63669 unique locations. (Tab: Visual_4)
2. Beat 506 has the highest number of homicide happened in all the records. (Tab: Visual_2)
3. In the neighborhood where over 85% of residents races are non-white, West End is the suburb has the most crime happened in all the records. (Tab: Visual_1)
4. Top 5 location types where has occurred the most crimes, consistently over the years were : House, Amenity, Shop, Building, Road. (Tab: Visual_6)
5. Quarter 3, has significant lower crime happened over the years, especially August, Where Larceny - Vehicle occurred the most frequent. (Tab: Visual_5)

Association Rules:

Please refer to the association rule tabs.

Startnet Model



StarNet Diagram

A StarNet Diagram was produced to support various business queries.

There are 7 dimensions in total as illustrated:

4 types dimensions including **TIME, CRIME, LOCATION TYPE, AND LOCATION**. To support any queries on any, or any combinations of these 4 types of dimensions.

For example :

- Which year, the location type club, has the highest numbers of crimes?
- Which suburb has least homicide crimes based on all the existed records?
- Which beat has the highest numbers of robbery crimes on the weekends in the first quarter of 2010 .

There are 4 different location dimensions. Due to the cross over classification / hierarchy issues.

For example :

- A certain main road may be stretched to different suburb.
- The **beat - zone** vs **neighborhood - npu** are two types of zoning method.

- More info can be found in the concept hierarchy pages.

ENTITY RELATIONSHIP DIAGRAM

A copy of ER diagram generated by SSMS.

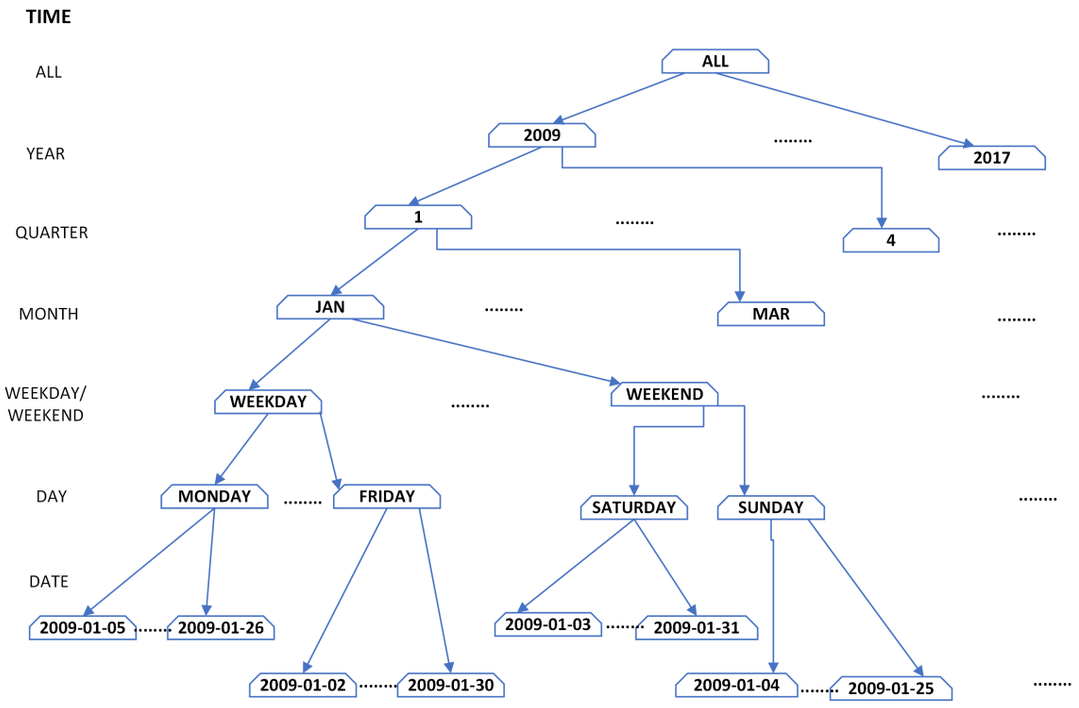
The star schema was implemented for the data warehouse design.

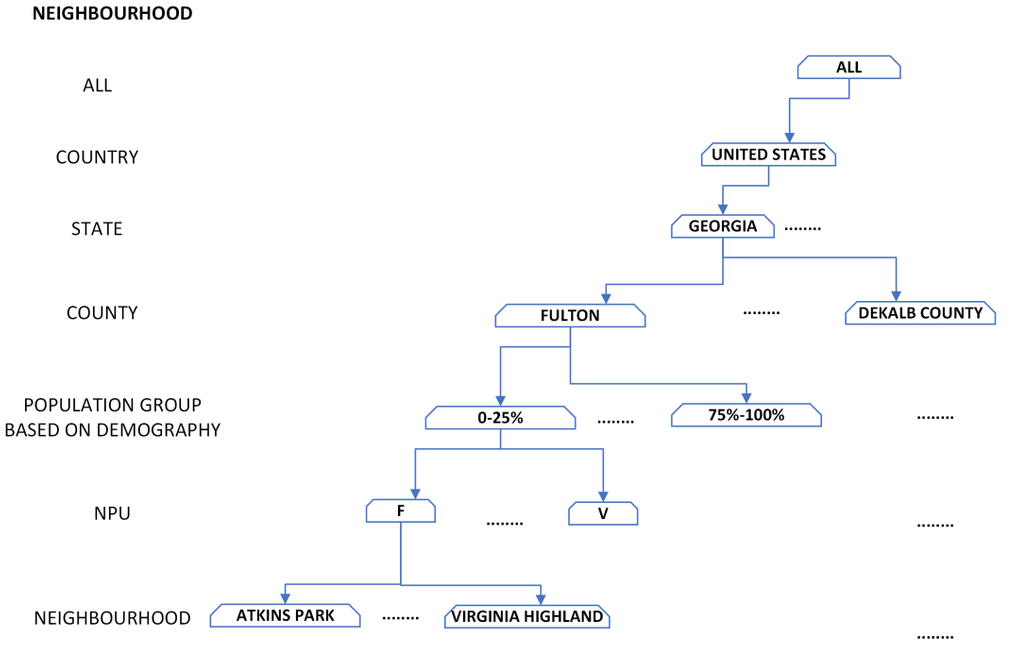
All the concept hierarchies for 7 dimentions tables were illustrated by using Microsoft Visio in the following two pages.

One 'factless' fact table was created, there was no measurement in the raw dataset, and no other measurement was applied.

The **COUNT** , which is the aggregation result of the number of crime occurrences, is the only measure in the cube design.

Concept Hierarchies for Each Dimension





TIME:

Through the ELT process, based on original date column, multiple hierarchies has been added to facilitate the queries.

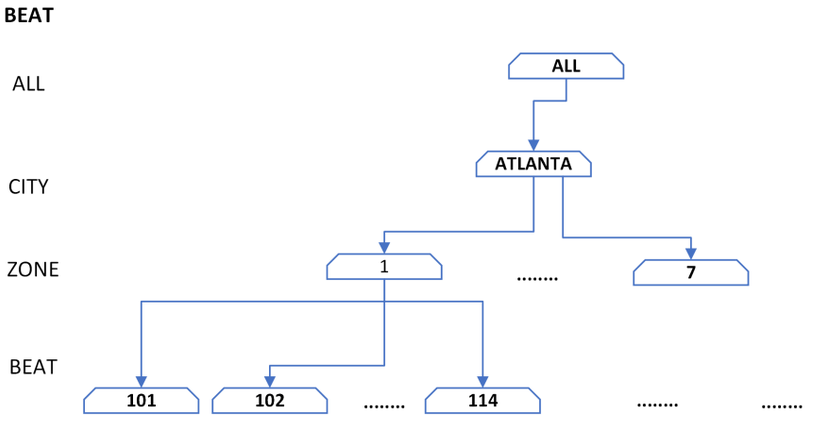
American format of date has been used.

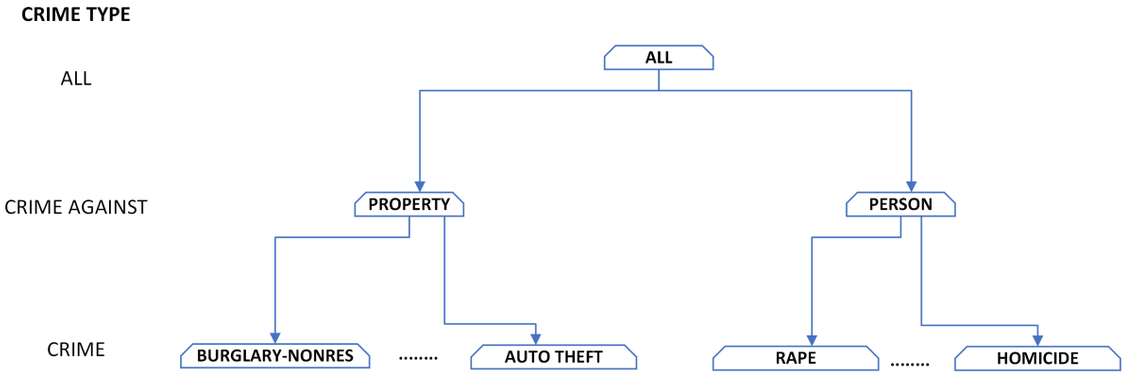
NEIGHBORHOOD:

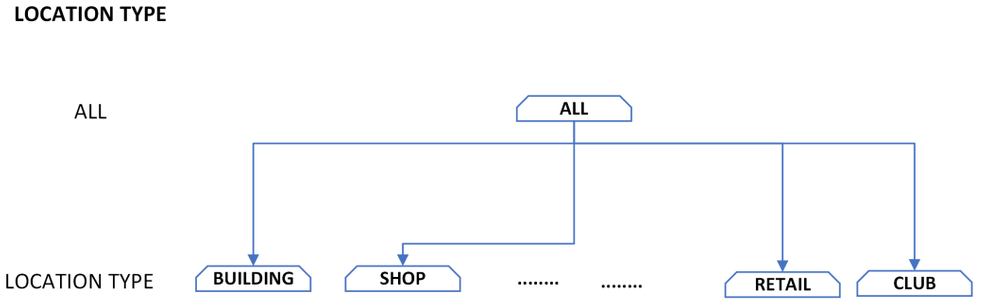
Through the ELT process, the data has been further enriched, based on original dataset coordinates columns, by using Python and Atlanta Gov City Planing ARCGIS map, JSON file was retrieved from ARCGIS rest server.

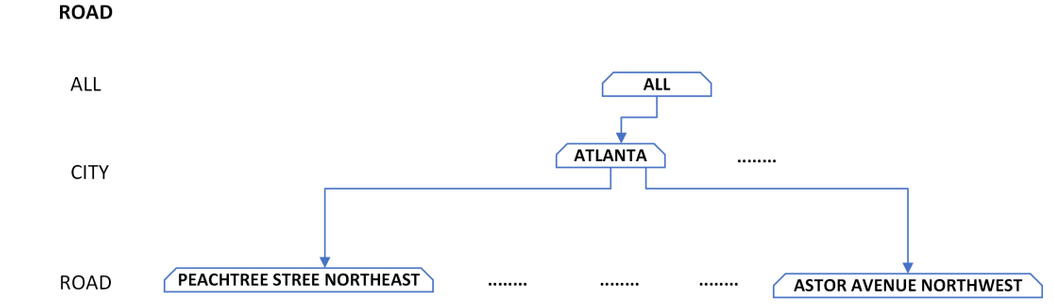
NPU: Neighborhood Planning Unit. More info

NPU was further grouped by demography/population data. Retrieved from Opendata Atlanta Statistics









BEAT :

Beat info was further been checked and enriched
Zone info has been added to create hierarchy.
Both was enriched by Python, by using JSON files retrieved from ARCGIS Rest Server.

Atlanta Interactive Map.
Beat JSON File Link Airport Beat JSON File Link
Zone Alignment JSON FILE LINK

CRIME TYPE:

A set-group hierarchy was created to group different crimes.

There are distinct difference between robbery, burglary and larceny. Based on whether the crime itself against property or person. Georgia Law Enforcement and Legislation. More info link

ROAD, LOCATION TYPE, COORDINATES:

* No extra hierarchies were introduced in these three dimension
* Lat/Long has been separated mainly for better PowerBI visualization, no hierarchy within

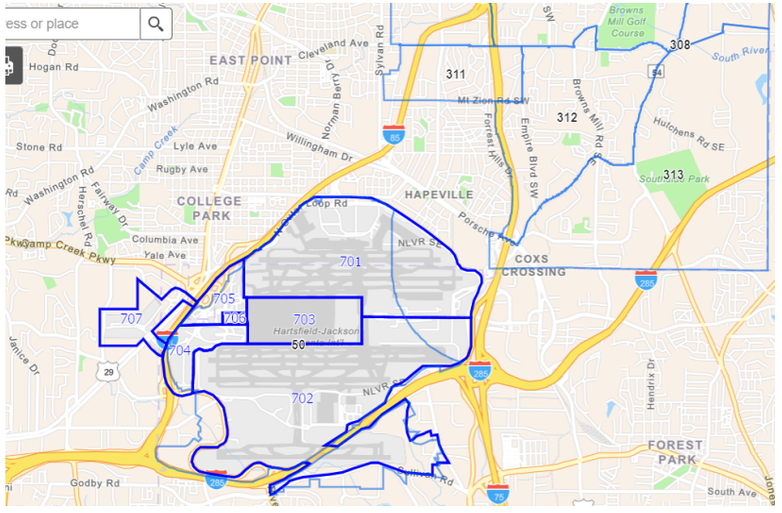
ETL

TOOLS: Python, SQL, MapShaper, ARCGIS Rest Server

PROCESS:

- **1:** Python Libraries : **pathlib, pandas, geopandas, glob, shapely.geometry, numpy, csv**
- **2:** Retrieve multiple raw data files, clean the format, and combine them all into one single dataframe, with the shape (**225001 rows × 17 columns**)
- **3:** Enrich Zone alignment info by using coordinates info column from original dataset, along with the JSON file from Atlanta ARCGIS Rest Server.
- **4:** Append Beat & Airport Beat alignment info into new column "Beat_2", by using coordinates info column from original dataset, along with the JSON files from Atlanta ARCGIS Rest Server.
- **5:** Due to discrepancies within the data, necessary steps and processes have been applied to the data. Especially the confusion between Zone 7 and Beat 700s, with Zone 50. The original dataset has lots of mix use of Beat 50 and Beat 700s. They are not identical area based on the map, however they are mostly overlapped. Based on the JSON file , the beat and airport beat layer. Zone 7 with Beat 700s has been decided to represent airport zoning. (sample shown below)

	crime	number	date	location	beat	neighborhood	npu	lat	long	type	road	neighbourhood_lookup	city	county
83358	LARCENY-FROM VEHICLE	121638018.0	06/11/2012	1255 SOUTH LOOP RD	702.0	NaN	NaN	33.62522	-84.43082	road	Airport Loop Road	NaN	NaN	Clayton County
109887	LARCENY-FROM VEHICLE	132028079.0	07/21/2013	1255 S LOOP RD	702.0	NaN	NaN	33.62522	-84.43082	road	Airport Loop Road	NaN	NaN	Clayton County
130472	LARCENY-NON VEHICLE	103128025.0	11/08/2010	1255 S LOOP RD @TECHNICAL SUPPORT CA	50.0	NaN	NaN	33.62522	-84.43082	road	Airport Loop Road	NaN	NaN	Clayton County
163920	LARCENY-FROM VEHICLE	93158035.0	11/11/2009	TECH CAMPUS @1255 S LOOP RD	50.0	NaN	NaN	33.62522	-84.43082	road	Airport Loop Road	NaN	NaN	Clayton County
193361	LARCENY-NON VEHICLE	100328060.0	02/01/2010	TECH CAMPUS @1255 S LOOP RD	50.0	NaN	NaN	33.62522	-84.43082	road	Airport Loop Road	NaN	NaN	Clayton County



- **6:** There are 125 rows of crime cases has no ZONE info, after being enriched by JSON. However the Zone info can be derived from the Beat info, Python function has been created to further finalized the missing zone info.
- **7:** Creating unique keys for each combination of the lat and long. In order to create Lat/Long dimension table.
- **8:** Checking city columns, there are 2039 out of 225,000 (less than 1%) data entry has no city info, and only one entry listed "Sandy Springs" as city. However upon checking ,the coordinate lies on the city boundary of Atlanta, so here the data has been included in our analysis for Atlanta crimes. Regarding all the nan values, the city Atlanta has been added in, even the airport zone is not in the city boundary, however most of the airport zone coordinates has Atlanta as city already, and the data set was about crimes in Atlanta.
- **9:** Creating unique keys for each beat + zone combination, then create beat dimension table which has beat - zone - city hierarchy, then exported to CSV.
- **10:** A date creator function has been created, in order to generate a date Dimension, and Date keys, which is the combination of date-month-year. The function automatically populate information for each hierarchy. Exported to CSV.
- **11:** Enrich neighborhood info from JSON file found on Atlanta Gov ARCGIS REST Server.
- **12:** Loading data from the Atlanta Official Statistic website, which has listed each neighborhood total population in 2017, and the population of each race composition. In order to have some demographic insight for each NPU (Neighborhood Planning Unit), a python function has been created, to combine, and group all the population figures of suburbs into NPU level groups, then calculation the percentage of each race, for each NPU.

  For the ease of set group hierarchy, the demography statistic has been group into the following 4 groups, across the NPU level:

  People of Color Population Percentage in each NPU.
  (0% ~ 25%], (25% ~ 50%], (50% ~ 75%], (75% ~ 100%]

```
stats_df. iloc[:, [1,2,3,4,7,8,9,10,11,12]]
```

	NPU	STATISTICA	POP2010	NEIGHBORHO	pop	white	black	asian	other	hispanic
0	NPU C	C04	2672	Arden/Habersham, Argonne Forest, Peachtree Bat...	2672	2573.136	24.048	26.720	13.360	34.736
1	NPU B	B10	3736	Peachtree Heights East, Peachtree Hills	3736	3235.376	220.424	74.720	67.248	138.232
2	NPU B	B01	4874	Peachtree Heights West	4874	3752.980	706.730	141.346	97.480	175.464
3	NPU B	B02	3372	Buckhead Forest, South Tuxedo Park	3372	2613.300	303.480	118.020	67.440	273.132
4	NPU A	A03	3423	Chastain Park, Tuxedo Park	3423	3210.774	23.961	88.998	37.653	61.614
...	...	...	...	...	...	...	...	...	...	...
97	NPU Z	Z02	2851	Thomasville Heights	2851	14.255	2788.278	5.702	31.361	14.255
98	NPU W	W02	3305	Benteen Park, Boulevard Heights, Custer/McDono...	3305	1054.295	1533.520	29.745	72.710	614.730
99	NPU W	W01	6827	Grant Park, Oakland	6827	4260.048	1904.733	150.194	170.675	341.350
100	NPU W	W03	4053	Ormewood Park	4053	2541.231	1146.999	93.219	101.325	166.173
101	NPU N	N01	3750	Cabbagetown, Reynoldstown	3750	2216.250	1162.500	86.250	105.000	180.000

	NPU	count	asian	black	hispanic	other	pop	white	NPU_POC_Pop_Pct
0	NPU A	3	3.48	3.01	2.22	1.28	100.0	90.05	(0%...25%]
1	NPU B	11	5.36	11.90	9.45	2.06	100.0	71.23	(25%...50%]
2	NPU C	6	3.23	8.26	6.04	1.48	100.0	81.00	(0%...25%]
3	NPU D	3	4.48	23.31	15.67	2.68	100.0	53.85	(25%...50%]
4	NPU E	7	12.61	17.03	4.88	3.04	100.0	62.49	(25%...50%]
5	NPU F	4	3.26	9.59	9.74	2.05	100.0	75.33	(0%...25%]
6	NPU G	3	0.43	93.53	1.88	1.34	100.0	2.82	(75%...100%]
7	NPU H	4	0.15	91.77	6.24	1.12	100.0	0.70	(75%...100%]

- **13:** Locate all the rows where newly enriched Neighborhood column is Nulll , but the original data's neighborhood info is present. then pass the info into new columns.
- **14:** Creating neighborhood dimension table, and unique keys. Export to CSV.
- **15:** Generate keys for each unique road record, export road dimension tables.
- **16:** Generate keys for each location type, export location type dimension table.
- **17:** Generate keys for each unique coordinate combination, export dimension table.
- **18:** Creating the same combination of the date as date keys for fact table, also as foreign keys to the date dimension table.
- **19:** Creating crime type dimension table, extra info added ,whether the crime is against person or property, in order to implement hierarchy. Export dimension table CSV.
- **20:** Create crime fact table, by using all the keys generated for each dimension table. Reindex the dataframe. Then finally exported to the crime fact table CSV.
- **21:** An issues was encountered due to the data type difference between JSON file keys and the PowerBI data type. Python was used to alter the Key data type for JSON file. In order to be imported as shape map for Power BI visualisation.
- **22:** https://mapshaper.org/ was used to convert **GeoJSON** file format to **TopoJSON**, for the custom shape map visualization in power BI

****All the above steps can be found in Python File, to notate each section of the codes. Most lines of the codes also have a ##comment line to explain the ETL process .**

	Unnamed: 0	crime	number	date	location	beat	neighborhood	npu	lat	long	type	road	neighbourhood_lookup
270678	82405	LARCENY-FROM VEHICLE	1.636620e+08	12/31/2016	855 PEACHTREE ST NE	505.0	Midtown	E	33.77798	-84.38389	NaN	NaN	NaN
270679	82406	LARCENY-FROM VEHICLE	1.636620e+08	12/31/2016	1241 EUCLID AVE NE	608.0	Candler Park	N	33.76676	-84.34724	NaN	NaN	NaN
270680	82407	AUTO THEFT	1.636622e+08	12/31/2016	2914 OLDKNOW DR NW	112.0	Collier Heights	I	33.77157	-84.48312	NaN	NaN	NaN
270681	82408	BURGLARY-RESIDENCE	1.636623e+08	12/31/2016	650 LYRIC WAY NW	112.0	Collier Heights	I	33.77304	-84.47748	NaN	NaN	NaN
270682	82409	BURGLARY-RESIDENCE	1.636623e+08	12/31/2016	144 MORELAND AVE NE	606.0	Reynoldstown	N	33.75698	-84.34929	NaN	NaN	NaN
270683	82410	LARCENY-FROM	1.636623e+08	12/31/2016	BAKER ST NE / PEACHTREE	508.0	Downtown	M	33.76226	-84.38755	NaN	NaN	NaN

	Blank	date_ID	coordinate_ID	beat_ID	Nbh_ID	Road_ID	Loc_type_ID	crime_ID
977	NaN	20091001	1	1	1	1	1	1
978	NaN	20091001	2	2	2	2	2	2
979	NaN	20091001	3	3	3	3	2	3
980	NaN	20091001	4	4	4	4	3	2
981	NaN	20091001	5	5	5	5	2	4
...	...	...	...	...	...	...	...	...
199525	NaN	20130202	4436	59	138	1109	6	8
199526	NaN	20130202	181	61	92	56	2	2
199527	NaN	20130202	18695	36	42	443	3	3
199528	NaN	20130202	8263	21	24	203	5	1
199529	NaN	20130202	10173	43	51	100	2	7

225000 rows × 8 columns

- **23:** SQL coding was used for database creation, key constraint, primary/foreign key relationship, bulk load data set into all the tables. Sample below. Please refer to the SQL code text document for details.

```
PRINT '';
PRINT '*** Dropping Database';
GO
IF EXISTS (SELECT [name] FROM [master].[sys].[databases]
WHERE [name] = N'DWCrime')
DROP DATABASE DWCrime;
GO
PRINT '';
PRINT '*** Creating Database';
GO
Create database DWCrime
Go
Use DWCrime
Go


PRINT '';
PRINT '*** Creating Table DimCrimeType';
GO
USE DWCrime
Create table DimCrimeType
(
CrimeID smallint primary key identity(1,1),
Crime varchar(50) not null,
CrimeAgainst varchar(10),
)
```

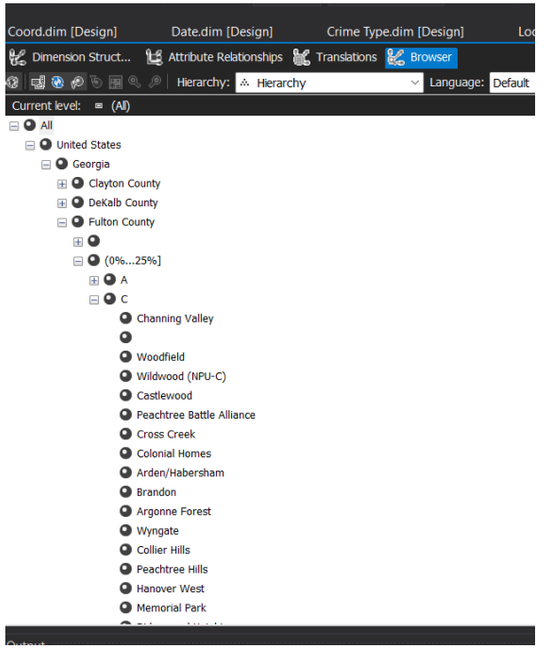
```
 BULK INSERT [dbo].[FactCrime] FROM 'C:\Users\coffe
\OneDrive\Desktop\Data Warehousing\Project 1\CrimeDW_Data
\crime_fact.csv'
 WITH (
    CHECK_CONSTRAINTS,
    --CODEPAGE='ACP',
    DATAFILETYPE='widechar',
    FIELDTERMINATOR='|',
    ROWTERMINATOR='\n',
    TABLOCK
);


ALTER TABLE FactCrime ADD CONSTRAINT
FK_Crime_ID FOREIGN KEY (Crime_ID) REFERENCES
DimCrimeType(CrimeID);
ALTER TABLE FactCrime ADD CONSTRAINT
FK_DateKey FOREIGN KEY (DateKey)REFERENCES
DimDate(DateKey);
ALTER TABLE FactCrime ADD CONSTRAINT
FK_Coordinate_ID FOREIGN KEY (Coordinate_ID)REFERENCES
DimCoord(Coordinate_ID);
ALTER TABLE FactCrime ADD CONSTRAINT
FK_Beat_ID FOREIGN KEY (Beat_ID)REFERENCES
DimBeat(Beat_ID);
ALTER TABLE FactCrime ADD CONSTRAINT
```

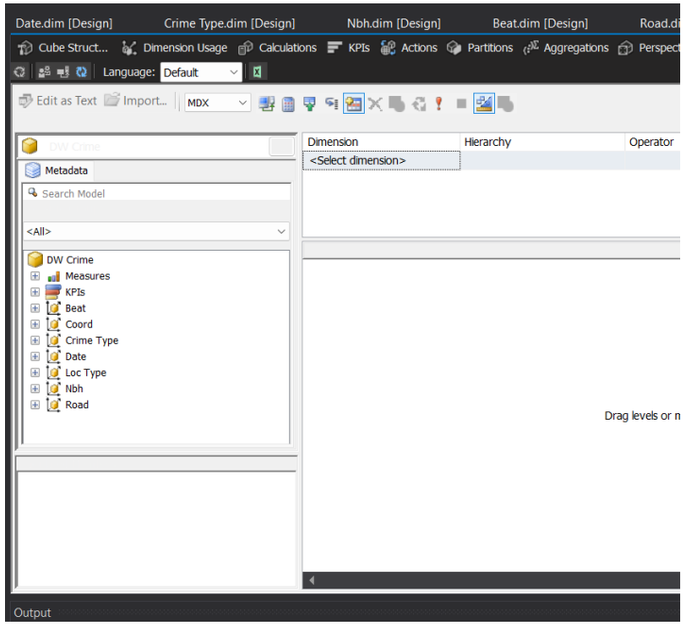
SSDT / SSAS / MOLAP / Cube Design

In order to analyze the data more quickly and efficiently, a multi-dimensional OLAP cube has been created, by using all the dimensions.

The total order concept hierarchies from all the dimensions were implemented in the cube. Please see below samples.

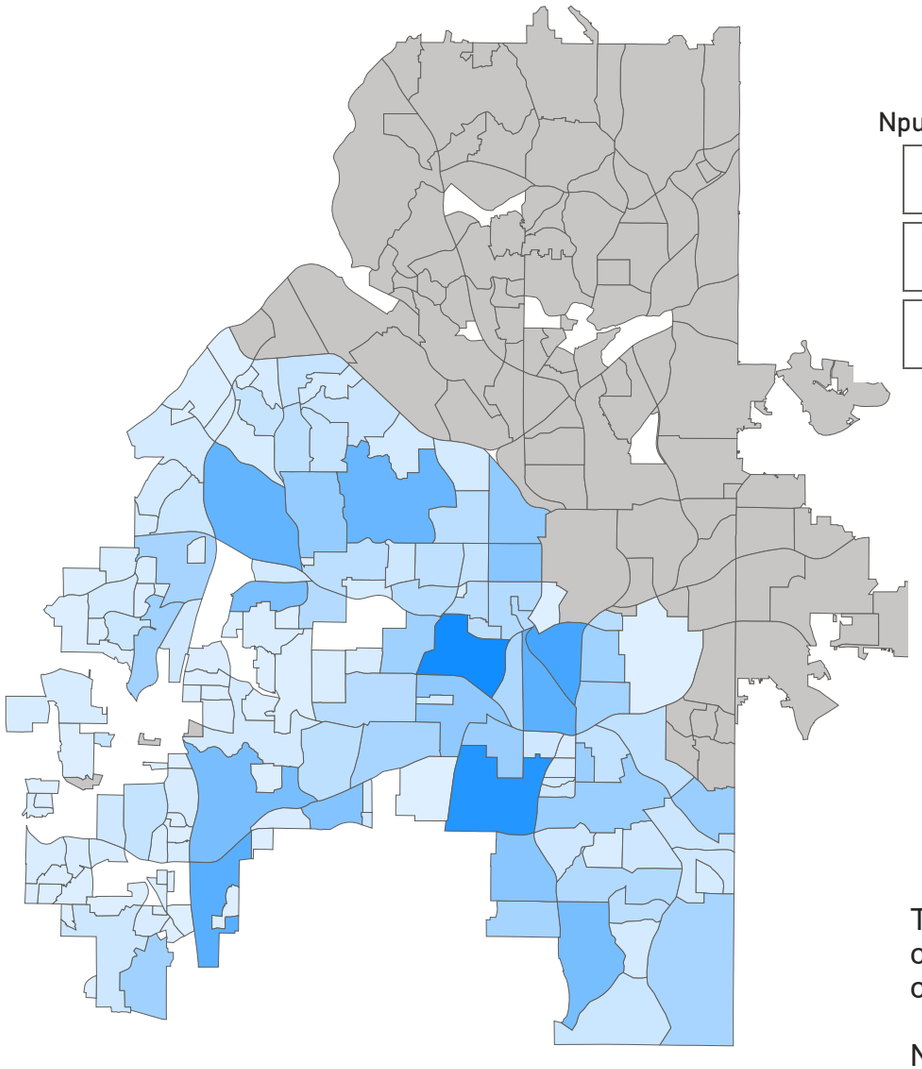
Please refer the solution project file of the SSDT analysis service multi-dimensional project for more details.





Power BI Visualization / Analysis

Crime Numbers by Neighborhood / NPU



Crime Type Slicer

Auto Theft

Year

All

Npu

G	J	P	S	X
H	K	Q	T	Y
I	L	R	V	Z

People of Color Population Percentage (NPU level): (75%...100%]

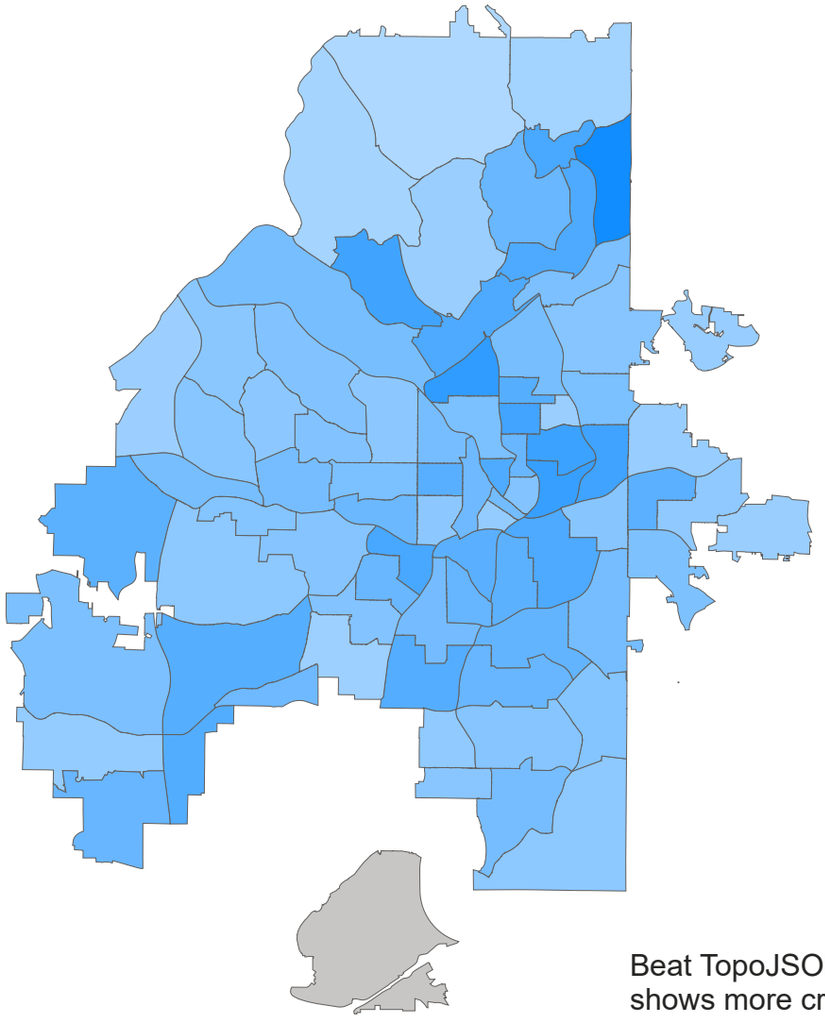
(0%...25%]	(50%...70%]
(25%...50%]	(75%...100%]

The neighborhood TopoJSON file has been loaded, color shades difference represents total crime number comparison, darker shades are the areas with more crimes, under the same criteria selection. Map area can be clicked as well.

NPU slicers are dynamically updated based on demography group slicer, along with year and crime type slicers, if you wish to look more specific aspect of historical crime records.

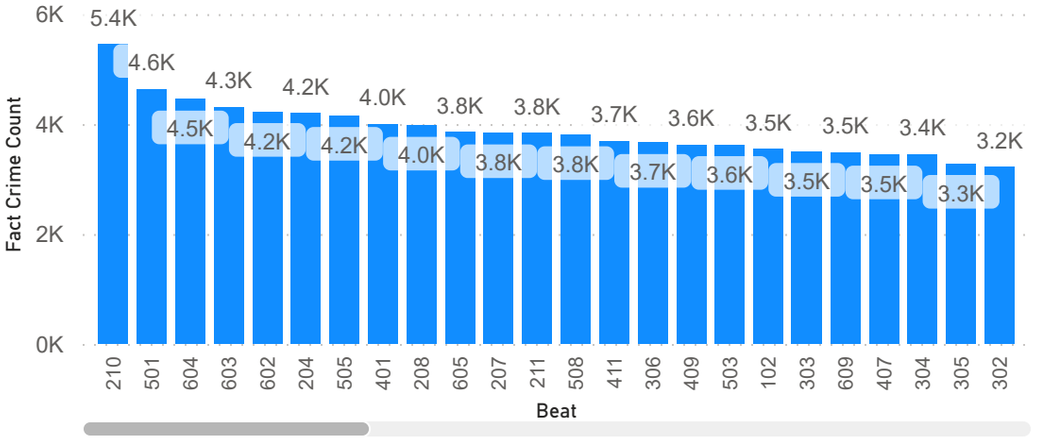
Please hover mouse over map for more details.

Crime Numbers by Beat / Zone



Crime Type

Select all	Robbery-Pedestri…	Larceny-From Vehicle
Burglary-Nonres	Burglary-Residence	Rape
Agg Assault	Robbery-Residence	Robbery-Comme…
Larceny-Non Vehicle	Auto Theft	Homicide



Zone

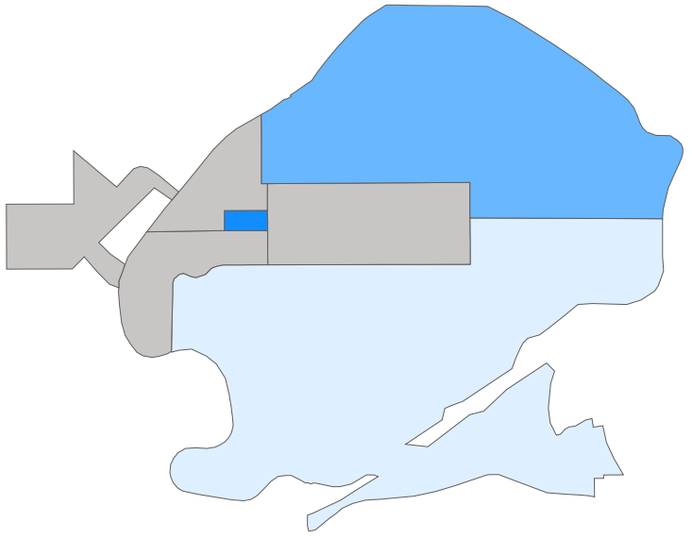
- [ ] 1
- [ ] 2
- [ ] 3
- [ ] 4
- [ ] 5
- [ ] 6
- [ ] 7

Beat TopoJSON files have been loaded for beat/zone visualization. The darker shades of color on the interactive map shows more crimes have happened, based on the selection criteria: Crime type, zone, etc.

The bar graph has been sorted in the order from the greatest number of crimes to the least number of crimes that has happened.

Map can be clicked on for any detailed exploration of certain areas. Please hover mouse over map for more details.

Crime Numbers by Beat / Zone

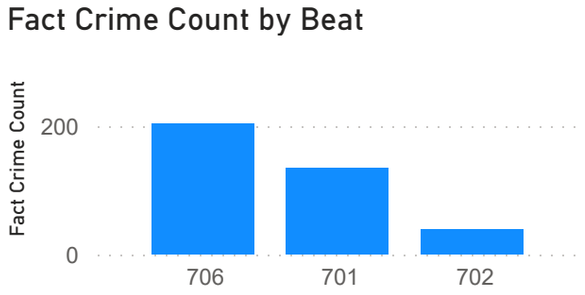


Beat

- 703
- 702
- 701
- 704
- 705
- 706
- 707

Crime Again...

- Persons
- Property



Due to the separate GeoJSON file between airport zone 7 and zone1-6, a follow up airport zone/beat visualization has been created in the same layout of previous slides.

However, more exploration to combine these two JSON files into one shape map will be needed for better information consolidation.

Overall, three TopoJSON files have been used:

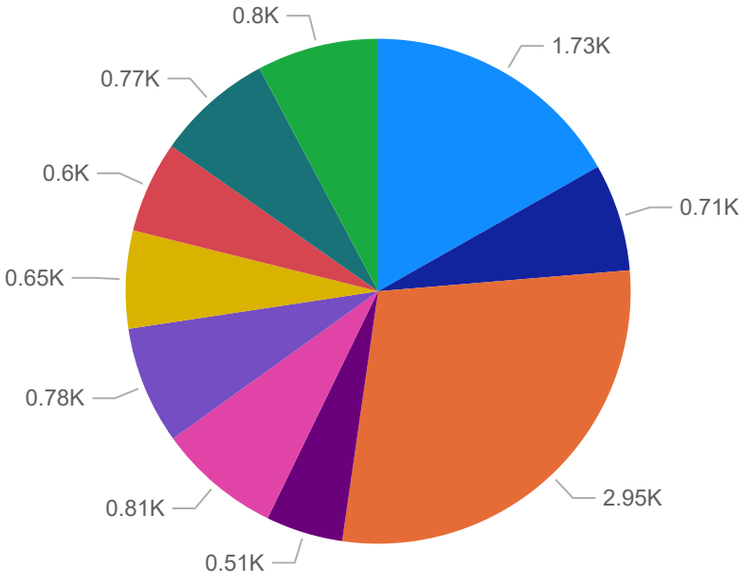
- **Neighborhood-NPU**
- **Beat - Zone**
- **Beat - Zone (Airport)**

The shape map visualization, can be extreme useful to interact with the data and communicate with stakeholders, for the better advancement of future city planning and police force arrangements/deployments.

All the visualization can be easily modified to better suit different type of enquiries, by adding time dimensions, location type dimensions, etc, Here only shows very few of all the different combinations.

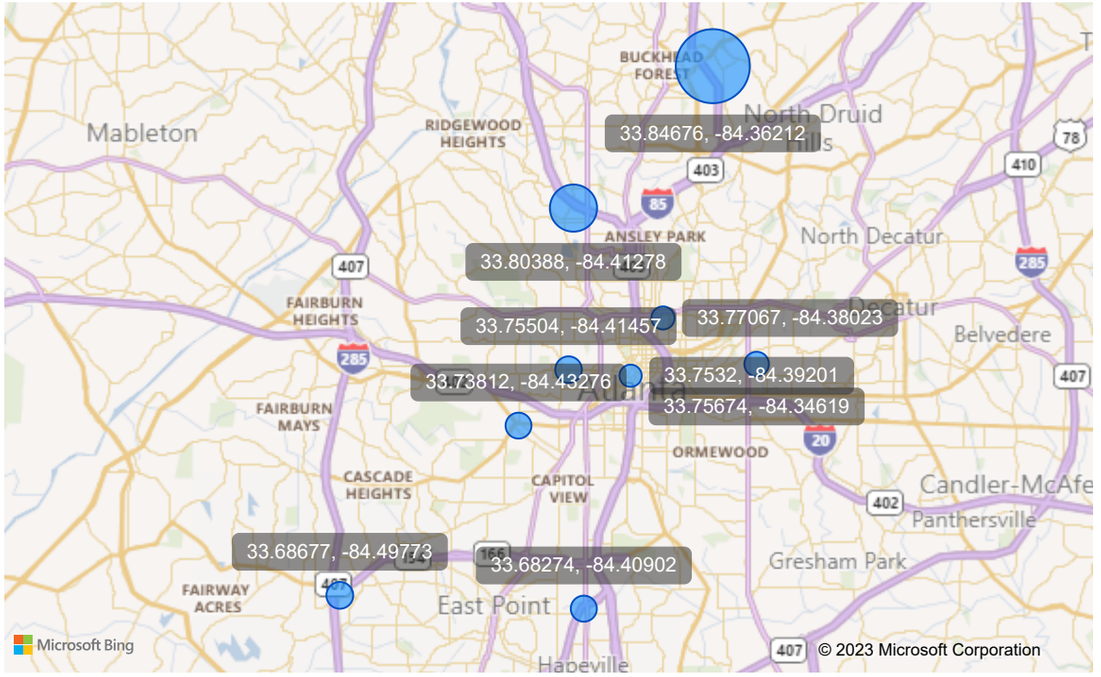
Top 10 locations(latitude, longitude), where the most crime occurrences has happened out of 225,000 crimes recorded for Atlanta from 2009 - 2017

***63669 unique latitude, longitude in total**



Coordinate

- (33.80388, -84.41278)
- (33.75674, -84.34619)
- (33.84676, -84.36212)
- (33.85302, -84.36278)
- (33.68677, -84.49773)
- (33.68274, -84.40902)
- (33.77067, -84.38023)
- (33.7532, -84.39201)
- (33.73812, -84.43276)
- (33.75504, -84.41457)



For the visualization purpose, a Lat/Long dimension table has been included in the SSDT, this may be subject to change / optimization due to no hierarchy within the dimension.

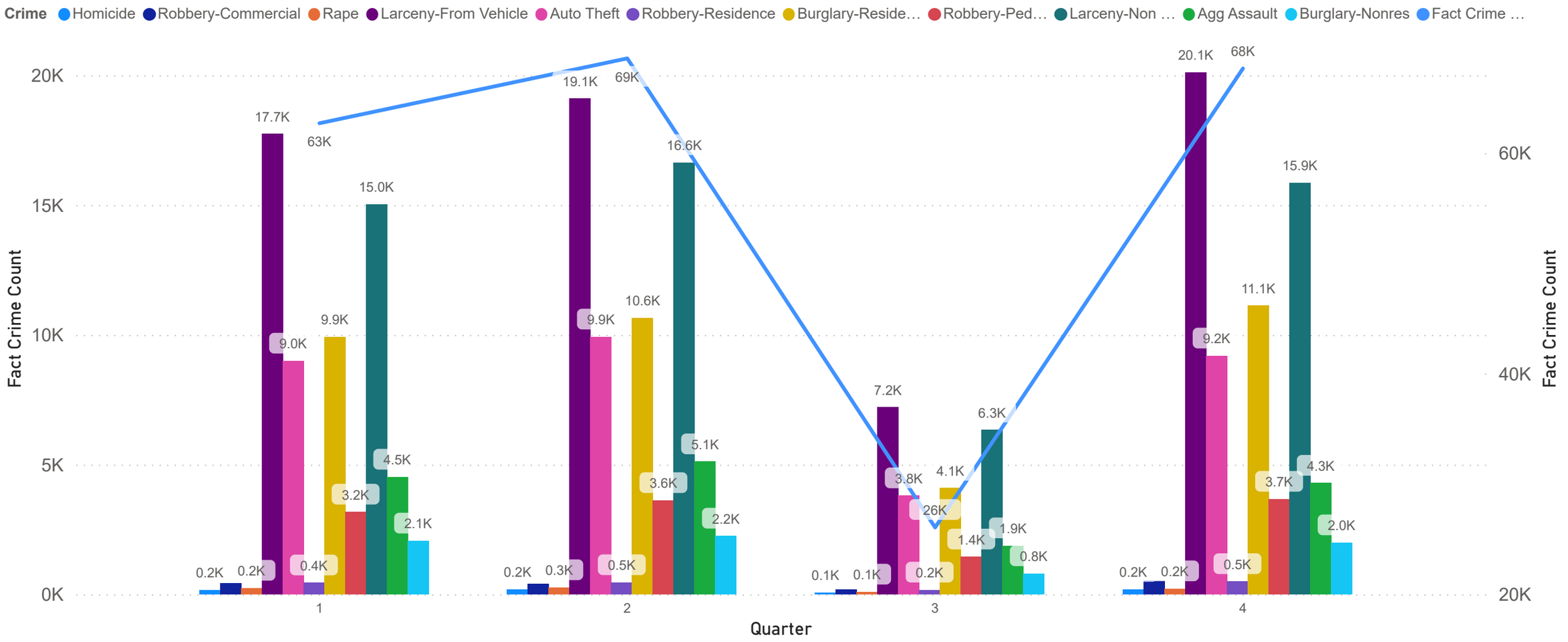
However, It is extremely interesting to discover :

Total Crime Records : 225,000
Total Unique Lat/Long Recorded : 63669
Average Crime Occurrence for Each Location: **4**

Location: (33.84676, -87.36212), the largest bubble size on the map
Crime Occurrence Recorded : **2946** , 1.31% of all the records.

****(33.84676, -87.36212): Lenox Square Shopping Mall . 3393 Peachtree Rd NE #3102, Atlanta, GA 30326, USA**

Roll up & Drill down Analysis - Date Dimension



Crime Against

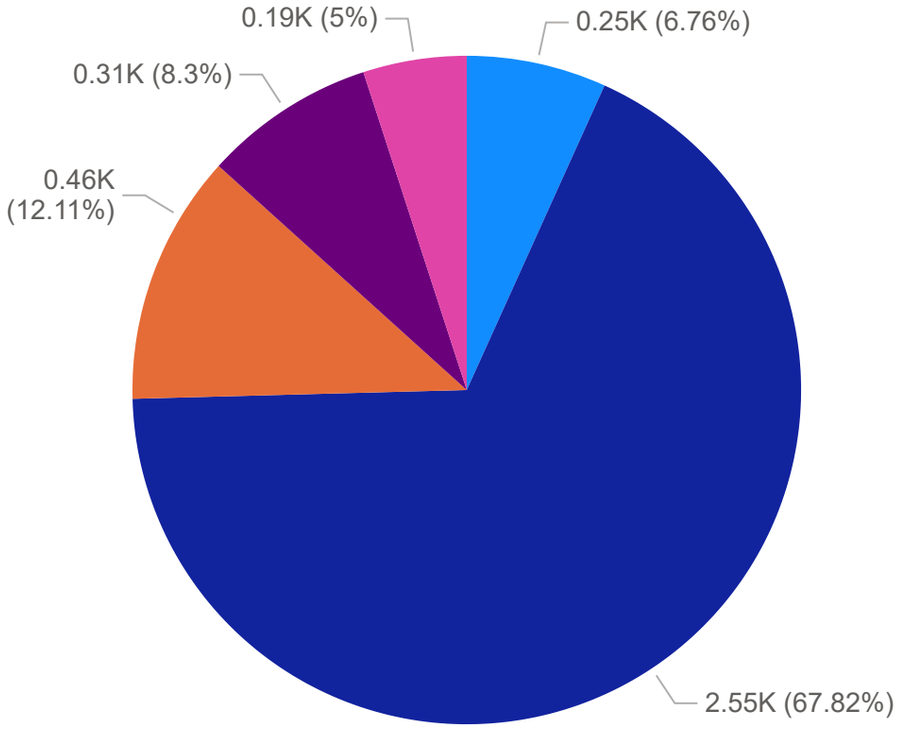
- [ ] Persons
- [ ] Property

Above visualization illustrated the Roll-up/ Drill down analysis, users will be able to navigate the icons on the top right corner, to visualizing the result on all the hierarchies in the time dimension, along with difference slicer selections for different aspects of the enquires. (only 1 slicer added here for cleaner presentation)

For the above particular case, Each bar chart result orders are in sync with the legend orders representing different crimes. The line represents the total crime numbers for each quarter. Multiple useful insights can be easily visualized, E.g.

- **Quarter 3** has the lowest crime counts, especially the **August** if the hierarchy is further drilled down onto month level.
- persistently **Larceny from Vehicle** claims the top number of crimes in all quarters. Homicide is the lowest. However this may change after drill down / roll up, or with slicer selections.

Top 5 location types occurred the most cirmes each year



Loc Type

- building
- house_number
- amenity
- shop
- road

Year

- [ ] 2009
- [ ] 2010
- [ ] 2011
- [ ] 2012
- [ ] 2013
- [ ] 2014
- [ ] 2015
- [ ] 2016
- [x] 2017

A simple visualization to explore what type of the location, had the crime happened the most frequently.

Association Rule Mining

For this part, the data has been further explored with association rule mining by using Visual Studio, please see below brief description / link for general concept.

Data Mining : a process of discovering actionable information from large set of data. link

Association Rules : association rules are "if-then" statements, that help to show the probability of relationships between data items, within large data sets in various types of databases. link

E.g. **{ A, B }** → **{ C }**

Support The percentage of transactions in all transactions that contain { A, B } & { C }

Confidence / Probability: The number of the transactions contain { A, B } & { C } **divided by** the number of the transactions contain only {A , B}

Lift / Importance : : Probability of { A, B } & { C } **divided by** (Probability of { A, B } times Probability of { C })

If Lift < 1 { A, B } & { C } are negatively associated
If Lift = 1 { A, B } & { C } are independent
If Lift > 1 { A, B } & { C } are positively associated

The goal of the data mining is to explore the association / relationship between different crimes. After multiple testing in order to produce meaningful association rules. I've decide to create two **views** as my **case table** and **nested table** for rule mining:

Case table :

Each unique combination of year - month and zone, has been treated as a unique transaction. In order to see what different crimes has happened within this criteria.

e.g. 2010 - June - Beat 103

```
CREATE VIEW view_Transaction_YM_BEAT AS
SELECT DISTINCT YM + CAST (Beat AS varchar(5)) as Month_Beat_Key, YM, Beat
FROM
(SELECT DISTINCT SUBSTRING ( CAST (dbo.FactCrime.Datekey AS VARCHAR(8)), 1, 6) as YM,
dbo.DimBeat.Beat, dbo.DimCrimeType.Crime
FROM    dbo.FactCrime INNER JOIN
            dbo.DimBeat ON dbo.FactCrime.Beat_ID = dbo.DimBeat.Beat_ID INNER JOIN
            dbo.DimCrimeType ON dbo.FactCrime.Crime_ID = dbo.DimCrimeType.CrimeID ) AS QUERY1
```

Results | Messages

	Month_Beat_Key	YM	Beat
1	200901101	200901	101
2	200901102	200901	102
3	200901103	200901	103
4	200901104	200901	104
5	200901105	200901	105
6	200901106	200901	106
7	200901107	200901	107

Nested table :

A table lists all the unique crime under each transaction.

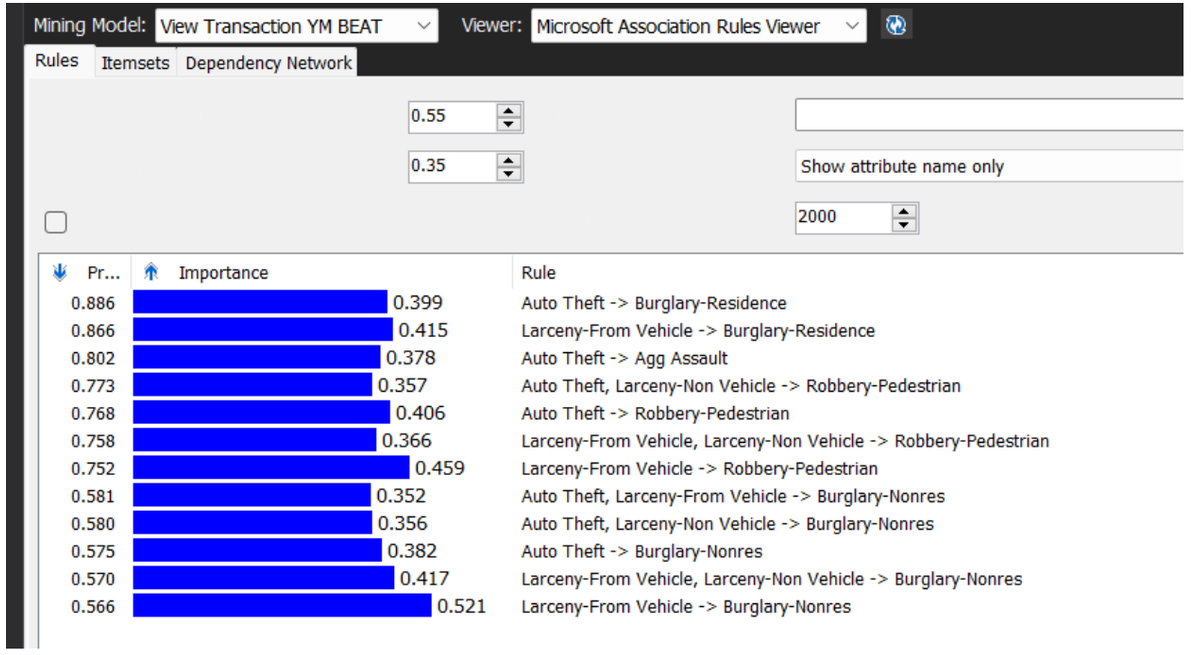
e.g. 2010 - June - Beat 103 Rape
2010 - June - Beat 103 Larceny - From Vehicle
2010 - June - Beat 104 Auto Theft

```
CREATE VIEW Crime_Type_YM_Beat AS
SELECT YM + CAST (Beat AS varchar(5)) as Month_Beat_Key, YM, Beat, Crime
FROM
(SELECT DISTINCT SUBSTRING ( CAST (dbo.FactCrime.Datekey AS VARCHAR(8)), 1, 6) as YM,
dbo.DimBeat.Beat, dbo.DimCrimeType.Crime
FROM    dbo.FactCrime INNER JOIN
            dbo.DimBeat ON dbo.FactCrime.Beat_ID = dbo.DimBeat.Beat_ID INNER JOIN
            dbo.DimCrimeType ON dbo.FactCrime.Crime_ID = dbo.DimCrimeType.CrimeID ) AS QUERY1
```

Results | Messages

	Month_Beat_Key	Crime
1	200901101	Auto Theft
2	200901101	Larceny-From Vehicle
3	200901101	Burglary-Residence
4	200901101	Agg Assault
5	200901101	Robbery-Residence
6	200901101	Larceny-Non Vehicle
7	200901101	Robbery-Pedestrian
8	200901101	Burglary-Nonres
9	200901102	Burglary-Residence
10	200901102	Robbery-Pedestrian
11	200901102	Auto Theft
12	200901102	Agg Assault
13	200901102	Larceny-Non Vehicle
14	200901102	Larceny-From Vehicle
15	200901103	Auto Theft
16	200901103	Robbery-Pedestrian

Association Rule Mining Result



Pr...	Importance	Rule
0.886	0.399	Auto Theft -> Burglary-Residence
0.866	0.415	Larceny-From Vehicle -> Burglary-Residence
0.802	0.378	Auto Theft -> Agg Assault
0.773	0.357	Auto Theft, Larceny-Non Vehicle -> Robbery-Pedestrian
0.768	0.406	Auto Theft -> Robbery-Pedestrian
0.758	0.366	Larceny-From Vehicle, Larceny-Non Vehicle -> Robbery-Pedestrian
0.752	0.459	Larceny-From Vehicle -> Robbery-Pedestrian
0.581	0.352	Auto Theft, Larceny-From Vehicle -> Burglary-Nonres
0.580	0.356	Auto Theft, Larceny-Non Vehicle -> Burglary-Nonres
0.575	0.382	Auto Theft -> Burglary-Nonres
0.570	0.417	Larceny-From Vehicle, Larceny-Non Vehicle -> Burglary-Nonres
0.566	0.521	Larceny-From Vehicle -> Burglary-Nonres

Rule View :

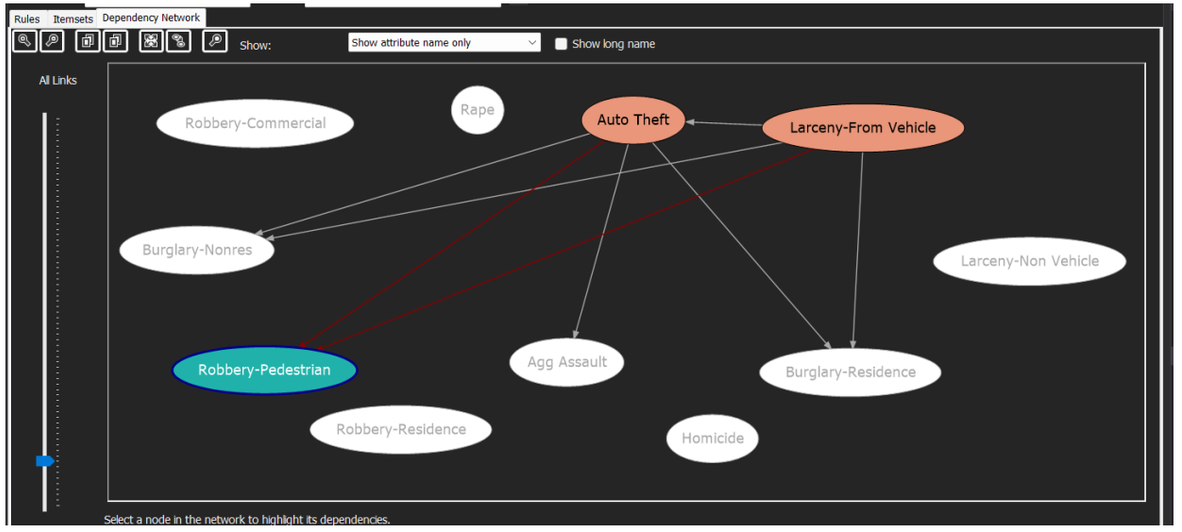
Minimum Probability : 55% Minimum Importance: 35%

The top 5 rules according to probability:

{ Auto Theft Burglary ----> Residence}
{ Larceny-From Vehicle ----> Burglary - Residence }
{ Auto Theft ----> Agg Assault }
{ Auto Theft, Larceny - Non Vehicle ----> Robbery - Pedestrian }
{ Auto Theft , Robbery - Pedestrian }

The top 5 rules according to importance:

{ Larceny-From Vehicle ----> Burglary-Nonres }
{ Larceny-From Vehicle ----> Robbery-Pedestrian }
{ Larceny-From Vehicle, Larceny-Non Vehicle ----> Burglary-Nonres }
{ Larceny-From Vehicle ----> Burglary-Residence }
{ Auto Theft ----> Robbery-Pedestrian }



Dependency Network View

The slider on the left has been lowered, in order to show the strongest link.
The arrow between nodes represents the association between crimes, the direction of the arrow dictates the prediction.

As shown on the left, Robbery-Pedestrian predicted by Auto Theft, Larceny-From Vehicle.

Further explanation on the rules discovered:

Probability	Importance	
0.570	**0.417**	**Larceny-From Vehicle, Larceny-Non Vehicle ----> Burglary-Nonres**

In all the transactions that contains **{ Larceny-From Vehicle, Larceny-Non Vehicle }, 57%** of these transactions also contains **{Burglary-Nonres} .**

Importance is 0.417, meaning the **{ Larceny-From Vehicle, Larceny-Non Vehicle } and {Burglary-Nonres} are negatively correlated**
The more occurrences of { Larceny-From Vehicle, Larceny-Non Vehicle } may means less {Burglary-Nonres} would happen.

Prediction based on this certain rules :

If the { Larceny-From Vehicle, Larceny-Non Vehicle } has happened, within that beat and that month:
There are 41.7% likelihood that {Burglary-Nonres} would also happen, or has happened.

Meanwhile, **{ Larceny-From Vehicle, Larceny-Non Vehicle } and {Burglary-Nonres} are negatively correlated**

** Please note, association doesn't mean causation, { Larceny-From Vehicle, Larceny-Non Vehicle } may, and may not, cause {Burglary-Nonres}, association rule mining does not provide insight on causation.

Discussion:

1 :

Different rules would be mined based on the design. What defines as "unique transaction" would directly affect the rule mining result.

If we treat each day, each neighborhood, as our item set criteria, each item set may have two few items for the mining.
If we treat each quarter, each zone, as our item set criteria, each item set may have too many items, meaning if most of the item set contains most of the crime types, then the mining process may also not be so viable.

2 :

The number of crimes was the main measure considered in this project. It would be ideal to combine the population figure for each level of location zoning, to have better comparison on how safe each suburb/beat/npu is. Less crime occurrences in a low population area may need more attention in some cases.
However the crime numbers results are still very useful for the police force deployment and government city planning.

3:

This project design can be further expanded with more data, more dimensions, more hierarchies, to suit various stakeholder needs. The visualizations can also be altered / customized under different requirements.