JANUARY 2021 | BUSINESS INTELLIGENCE COURSE

Theft: What did happen at Chicago from 2012 to 2017?

VNU – Ho Chi Minh University of Science
Nguyễn Duy Cường | ndcuong869@gmail.com

Table of Contents

I. Business understanding

This report represents how to build a complete solution for data analysis and mining project with existed support tools include SSIS, SSAS, Power BI and Python. It is a tutorial help the beginner understand stages and which problems should be considered when build a data warehouse. Also, it explains technical skills to improve performance without very hard coding.

Crime is a one of major issues that all countries have been facing recently. Analysis crimes behaviors is an important step to the government identify and provide set of ideas to reduce crime rate in the country. This report only concentrates on analyzing and mining theft crimes which is classified by IUCR organization.

Primary objectives

- Design and implement an end-to-end data warehouse.
- Use tools to make report, OLAP and data mining from the data warehouse.
- Assess and remark the overall system.

Secondary objectives

- Analysis theft behaviors through community area, location, time-series and domestic.
- Determine which factors can affect to theft behaviors to provide an insight to violent activities.
- Build a supervised model to predict a case can be classified to whether crime or not.

II. Data understanding

This stage should be completed with a set of datasets can fulfill business requirements. It should determine which features is used to accomplish objectives.

To achieve defined goals, a data warehouse should be built and structured associated with above requirements. Please view appendix A to understand features for each dataset.

Original datasets

The Chicago crimes

The data sources can be downloaded at this link.

Link: https://www.kaggle.com/currie32/crimes-in-chicago

External datasets

External datasets can explore more information and insight for this project.

FBI Code

Link: https://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html

Represent classification type and definition for each crime type.

Because the original data source did not represent as a table structure, so it needs re-structure to be analyzed by using this script.

Socioeconomic

Link: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

Contains a selection of six socioeconomic indicator of public health significance by Chicago community areas.

III. Data preparation

This stage is constructed to clean data by identify missing, duplicates and outlier values. Furthermore, it can wrangle data from many sources.

Data preparation can use many tools such as Pentaho, OpenRefine, IBM InforSphere, …

In this report, it is used by Python script with many modules support for data cleaning.

Handle missing values in datasets

Chicago Crimes dataset has missing values is showed in below table.

Column	NULL count	Handle method
Location Description	1658	Keep them as null value
Community Area	53	Have 40 missing values and 13 zero values. Because community area begins at 1. So, zero value is a missing value. Change all to zero value.
X Coordinate	37083	Drop this column
Y Coordinate	37083	Drop this column
Latitude	37083	Drop this column
Longitude	37083	Drop this column
Location	37083	Drop this column

From X Coordinate and Y Coordinate can inference to Latitude, Longitude and Location. It is reason why they same null count for each column. Furthermore, the column location description represents information same location, so it can be dropped out of dataset without affect analysis process.

Determine and handle duplicates data

Column Location contain some duplicates and wrong values such as ‘TaxiCab’ and ‘Taxi Cab’, ‘PoolRoom’ and ‘Pool Room’, … To handle this problem, a data dictionary is used to find and replace strategy.

Key	Value
TAXICAB	TAXI CAB
"CTA ""L"" PLATFORM"	CTA PLATFORM
"CTA ""L"" TRAIN"	CTA TRAIN
MOTEL	HOTEL/MOTEL
HOTEL	HOTEL/MOTEL
NURSING HOME	NURSING HOME/RETIREMENT HOME
POOLROOM	POOL ROOM
HALLWAY	RESIDENCE PORCH/HALLWAY
VACANT LOT	VACANT LOT/LAND
TAVERN	TAVERN/LIQUOR STORE
GOVERNMENT BUILDING	GOVERNMENT BUILDING/PROPERTY
GARAGE	GARAGE/AUTO REPAIR
BARBERSHOP	BARBER SHOP/BEAUTY SALON
NULL	UNKNOWN

IV. Data modeling

This section represents how Chicago crimes data is structured from stage, normalization data store (NDS) and dimension data store (DDS). It covers a schema, transforming and meaning in design.

SQL Server database is used to implement data stores in this step because this framework has variety of components support ETL steps and integrate data from many sources. Furthermore, it can be executed automatically by the administration.

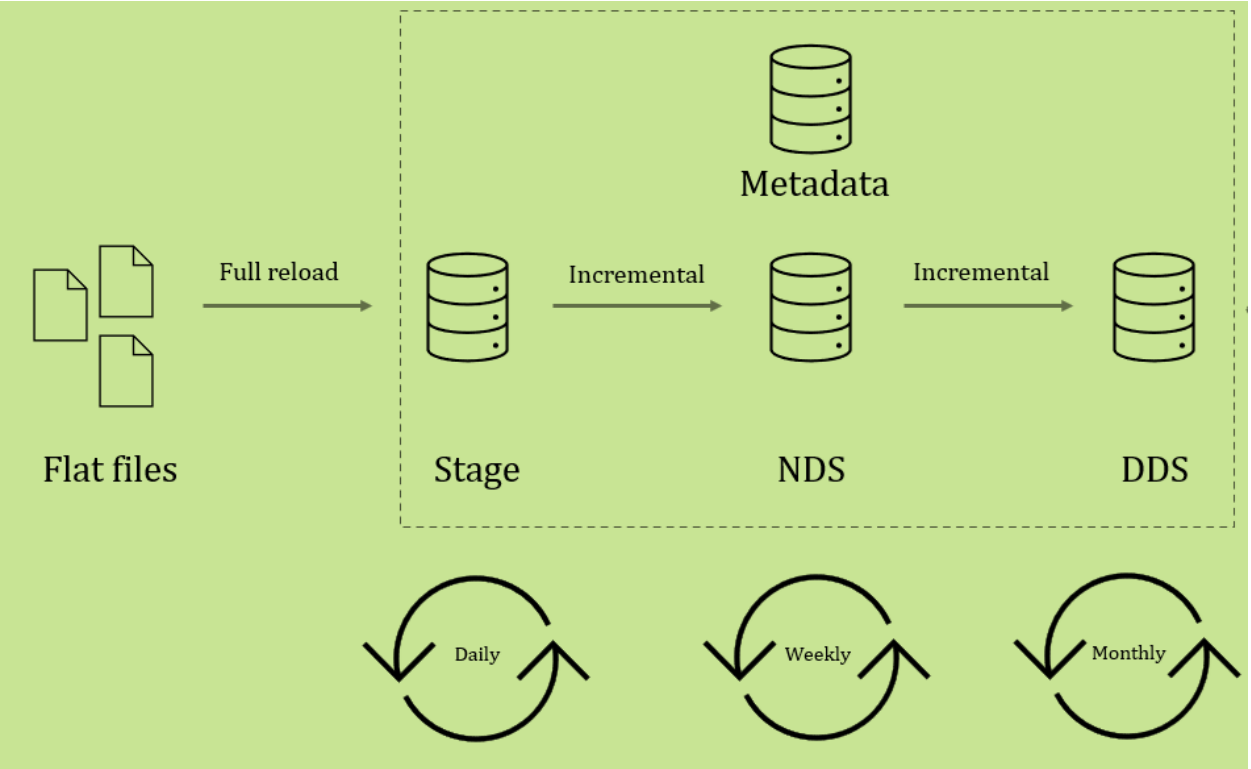

Figure 1 Data warehouse architecture

Stage data store design

Stage stored data is collected from sources, so tables should keep information same with data sources. To get more information about tables include data type and description, please view at APPENDIX A.

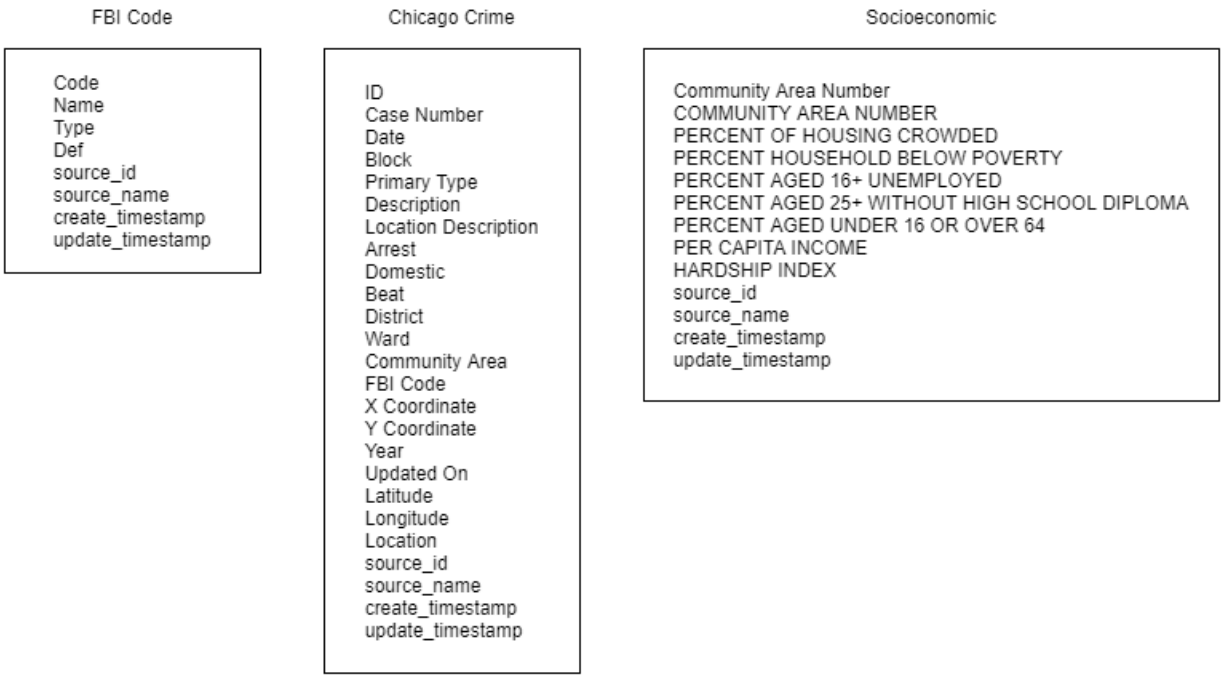

Figure 2 Stage design

Some notes when design and implement stage data store:

- Should not create integration constraint to capture data quality for report.
- Add source features include source id and source name.
- Add time features include create and update time.

Because load data from data source to stage using full reload, tables data should be truncated before running packages.

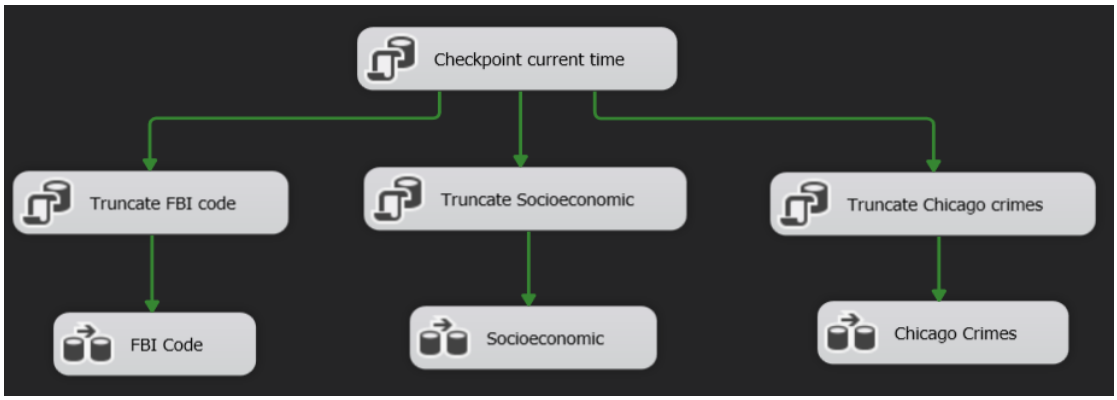

Figure 3 Stage package in SSIS

Control	Description
Checkpoint current time	Save execution time to metadata
Truncate FBI code	Truncate FBI table in stage
Truncate Socioeconomic	Truncate Socioeconomic table in stage
Truncate Chicago crimes	Truncate Chicago Crime table in stage
FBI Code	Load FBI Code data from flat file to stage
Socioeconomic	Load Socioeconomic data from flat file to stage
Chicago Crimes	Load Chicago data from flat file to stage

Figure 4 Control description in stage packages

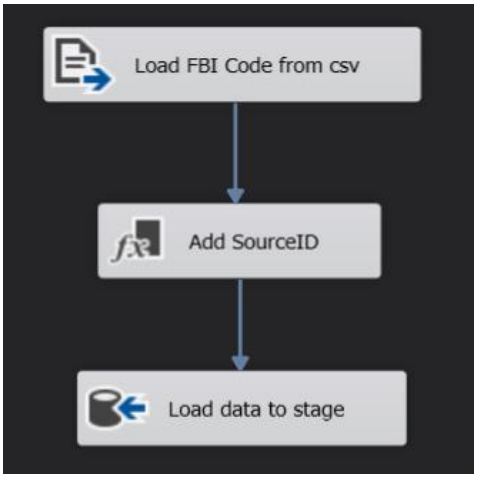

Figure 5 FBI Code data flow

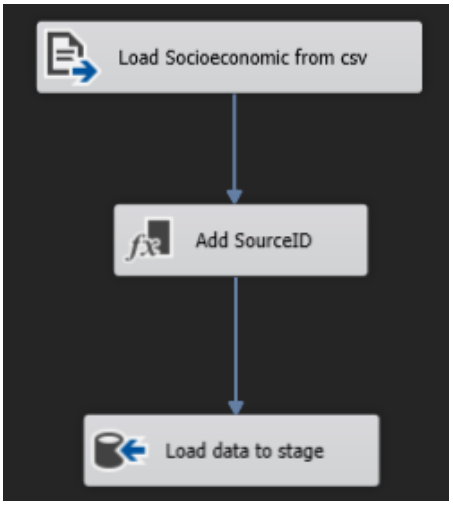

Figure 6 Socioeconomic data flow

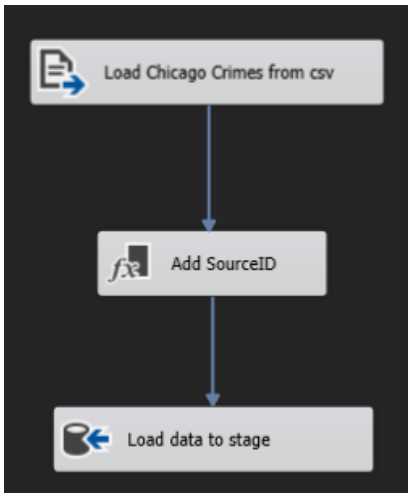

Figure 7 Chicago Crime data flow

Normalization data store design

To get more information about individual feature transforming from stage to NDS, please view at Appendix B

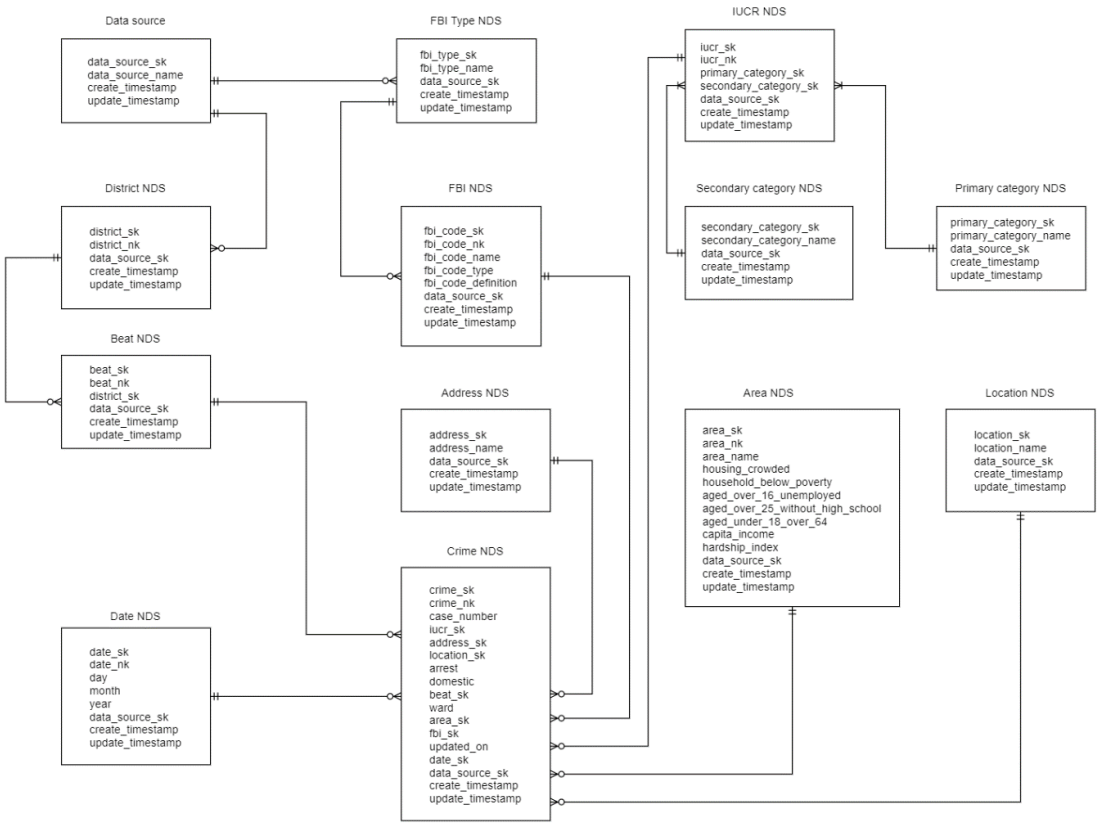

Figure 8 NDS Design

Notes

When design NDS for data analysis and mining project, all NDS tables should contain unknown values because an instance should reference the other instead of leaving null value.

Figure 9 NDS package in SSIS

Control	Description
Get NDS last updated	Get last updated time to incremental loading
Checkpoint execution time	Save current time to metadata
Data Source	Populate data source from stage to NDS
Community Area	Populate community area from stage to NDS
FBI Code Type	Populate FBI type from stage to NDS
FBI Code	Populate FBI code from stage to NDS
Primary category	Populate primary category from stage to NDS
Secondary category	Populate secondary category from stage to NDS
IUCR	Populate IUCR from stage to NDS
Address	Populate address from stage to NDS
Location	Populate location from stage to NDS

District	Populate district from stage to NDS
Beat	Populate beat from stage to NDS
Date	Populate date from stage to NDS
Crimes	Populate crime from stage to NDS
Checkpoint update time	Save execution package to last update data

Figure 10 Control description in NDS package

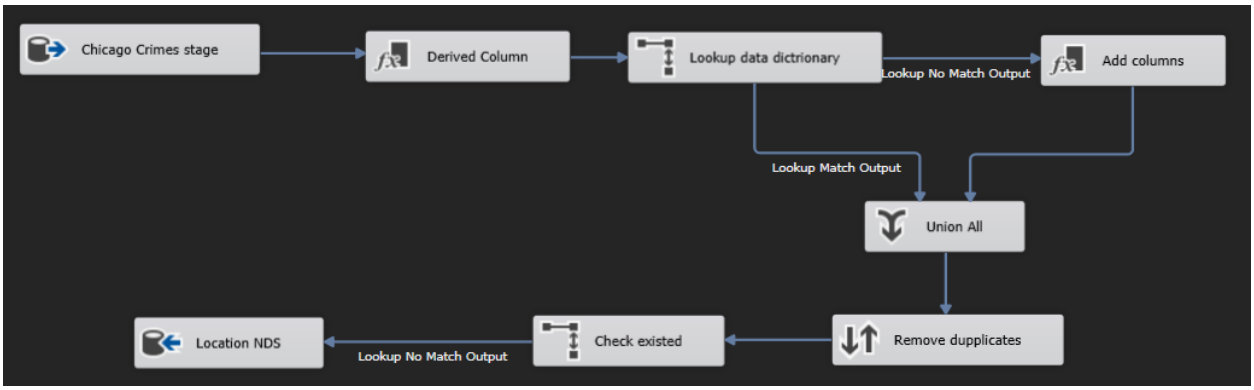

Figure 11 Location NDS using data dictionary

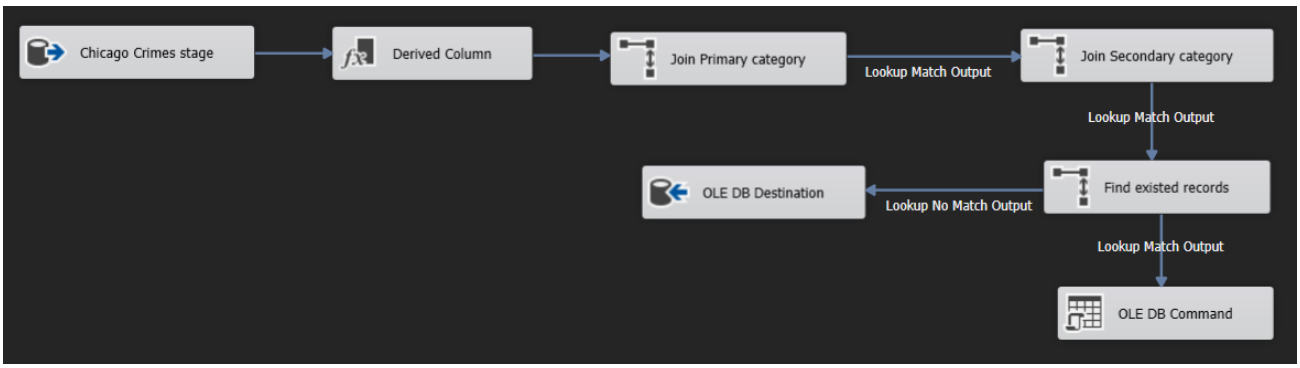

Figure 12 IUCR NDS data flow

Dimension data store design

To get more information about individual feature transforming from NDS to DDS, please view Appendix C

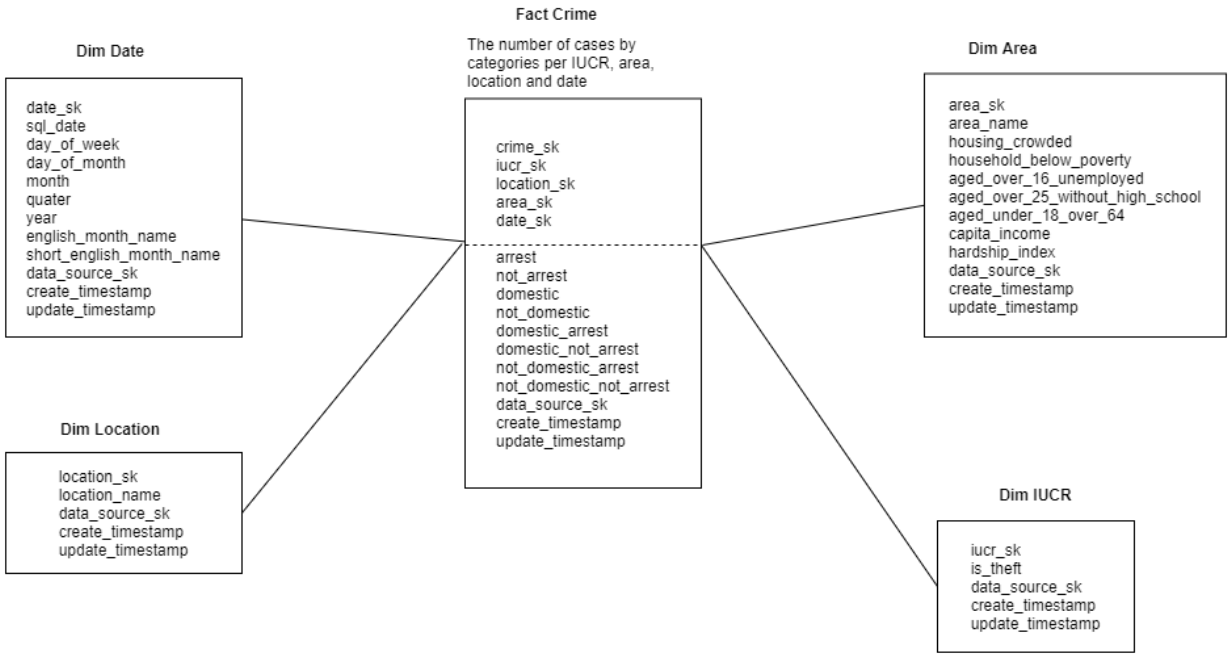

Figure 13 DDS design

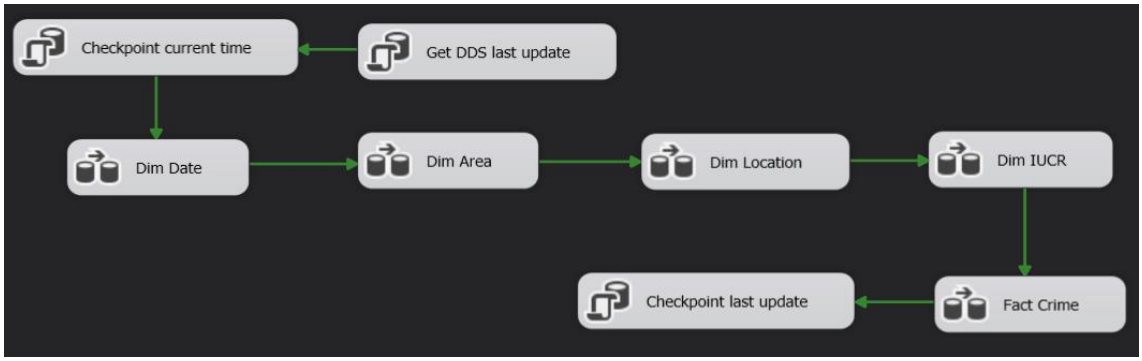

Figure 14 DDS package in SSIS

Control	Description
Get DDS last update	Get last update time in metadata
Checkpoint current time	Save current time to metadata
Dim Date	Populate date from NDS to DDS
Dim Area	Populate area from NDS to DDS
Dim Location	Populate location from NDS to DDS
Dim IUCR	Populate IUCR from NDS to DDS
Fact Crime	Populate crime from NDS to DDS
Checkpoint last update	Save execution package to metadata

Figure 15 Control description in DDS package

Populate data to DDS take a lot of time to execute package in the first time because the Slowly change dimension (SCD) component operate comparing between records. To get a better performance, SCD should be used in the next time instead of first executing.

To improve overall system performance, view Appendix D – Best practice in SSIS.

Packages Scheduling

SQL Server has installed Integration Service to create a job which can automatically run without explicit operations. After a package is executed, it will send a mail to DBAs.

Package name	Scheduler
Stage daily full reload	23:55 every day
NDS weekly incremental	23:55 Sunday
DDS monthly incremental	23:55 last day of month

Figure 16 Package scheduler on SQL Server

The email notification looks like below image.

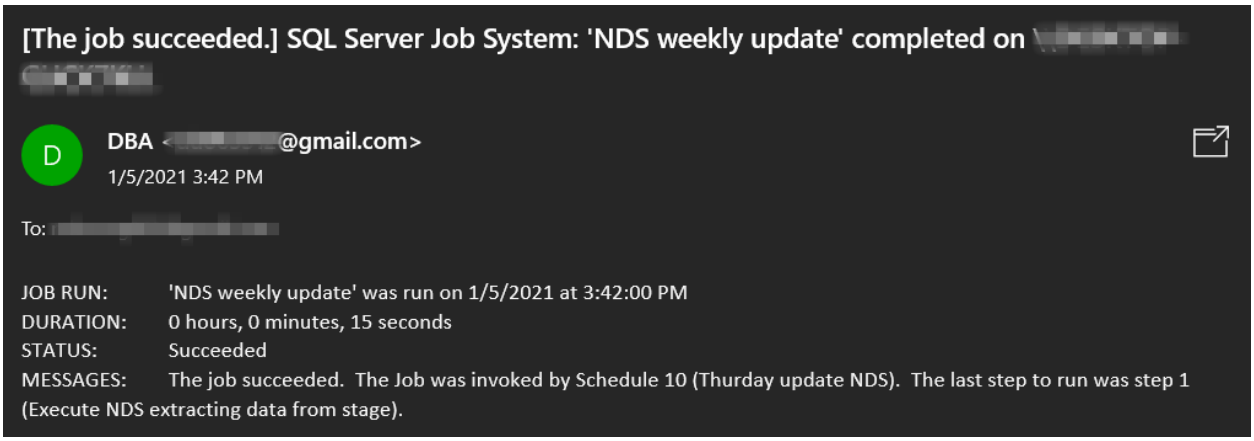

Figure 17 SQL Server job notification

V. Report and mining data

After implementing the data warehouse, data should be mined to provide insight about business requirements. There are many approaches to analysis data, however, this section only presents 3 methods include reporting, OLAP and data mining.

Approach name	Description	How to
Reporting	A simple report retrieves a few columns to present them in tabular format on the screen	Connect directly to DDS to get data and create a report
OLAP	Enable the business users can go up/drill down to a particular area of MDB to view data at a higher/more detail levels	Access data by using a cube which is built on DDS
Data mining	Discovering the pattern in data	Access data directly from DDS to build a model

Figure 18 Data analysis methods

Reporting

This activity show what happened in data by many forms such as tabular, crosstab, charts, … To provide an insight clearly, crime data should be analyzed by using Power BI application. Design files locate on the Power BI folder.

In this section, theft behaviors are analyzed by location, time-series, community area by using visualization and statistic.

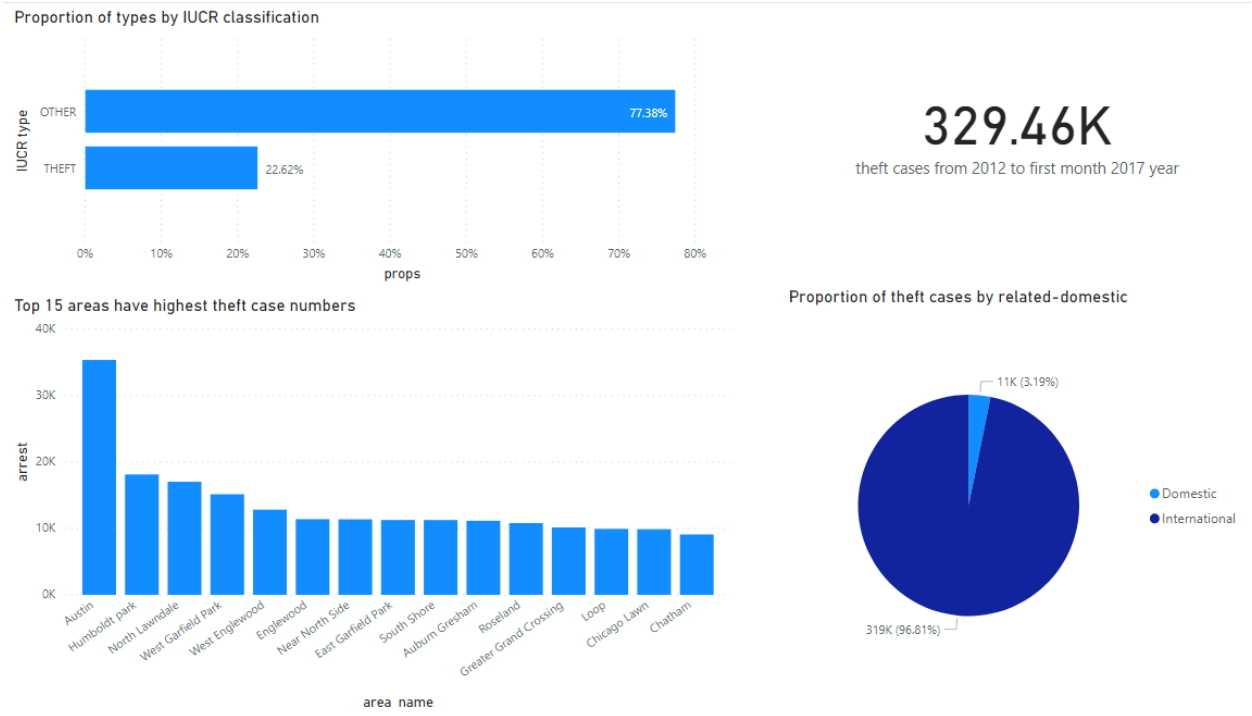

Figure 19 Overview of theft crimes at Chicago country. Proportion of types by IUCR classification (top left). Top 15 areas have highest theft case rate (bottom left). Number of theft cases from 2012 to 2017 (top right). Proportion of theft cases by related-domestic (bottom right).

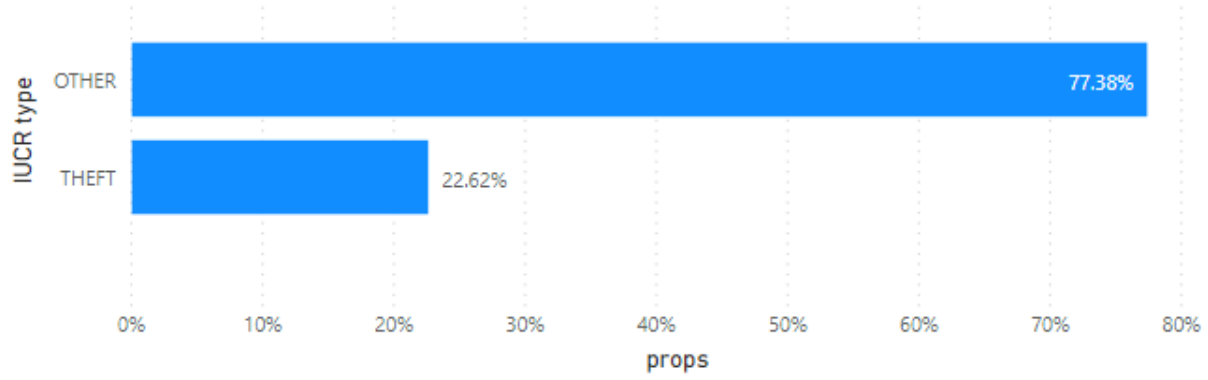

Figure 20 Proportion of types by IUCR classification

Proportion of theft crimes in total numbers nearly one quarter with 22.62%. Assumption that the robbery is popular at Chicago country.

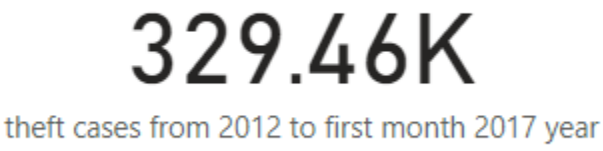

Figure 21 Total number of theft cases from 2012 to first month 2017 year

There were approximate 330K robberies for 5 years. This is a huge number to believe that theft is a critical issue in the Chicago social.

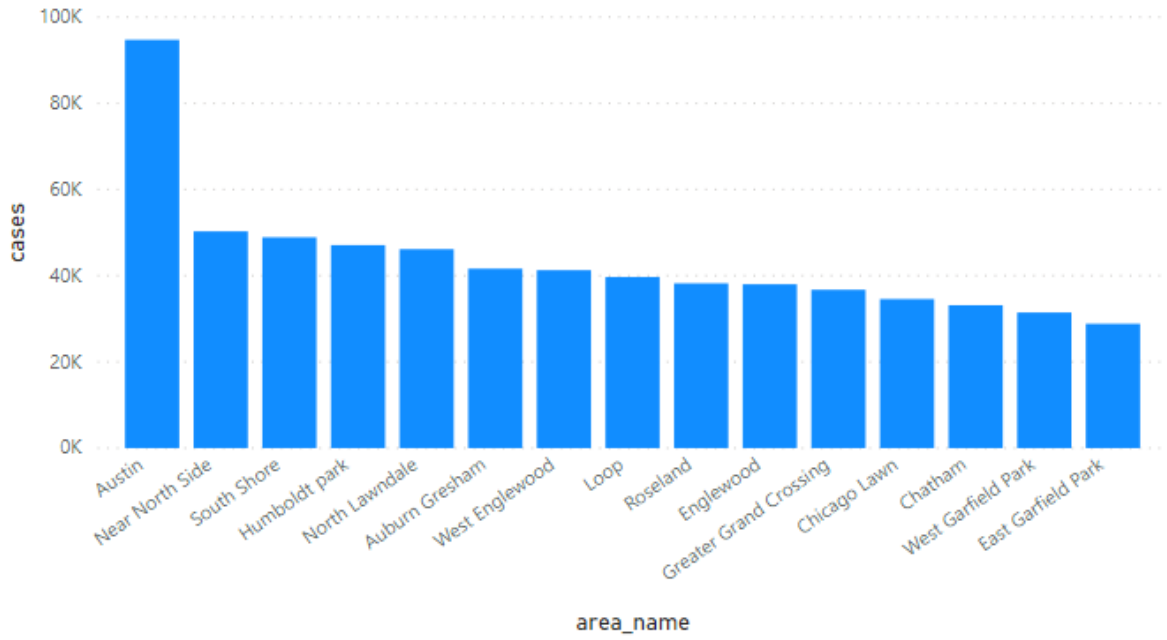

Figure 22 Top 15 areas have highest theft case numbers.

Some cities have higher theft crime rate than others with Austin city saw roughly 100k incidences of burglaries. They should be analyzed properties to understand why those cities usually noticed as a dangerous area.

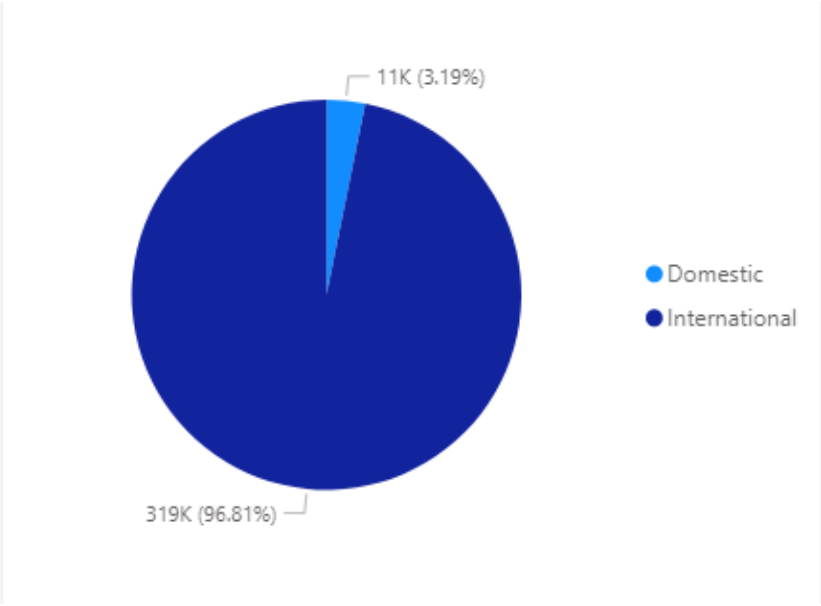

Figure 23 Proportion of theft case numbers by related-domestic.

The pie chart show that international crimes had higher crimes rate compared to domestic with around 97%. This is explained that Chicago country has a high non-native rate which have lived in here.

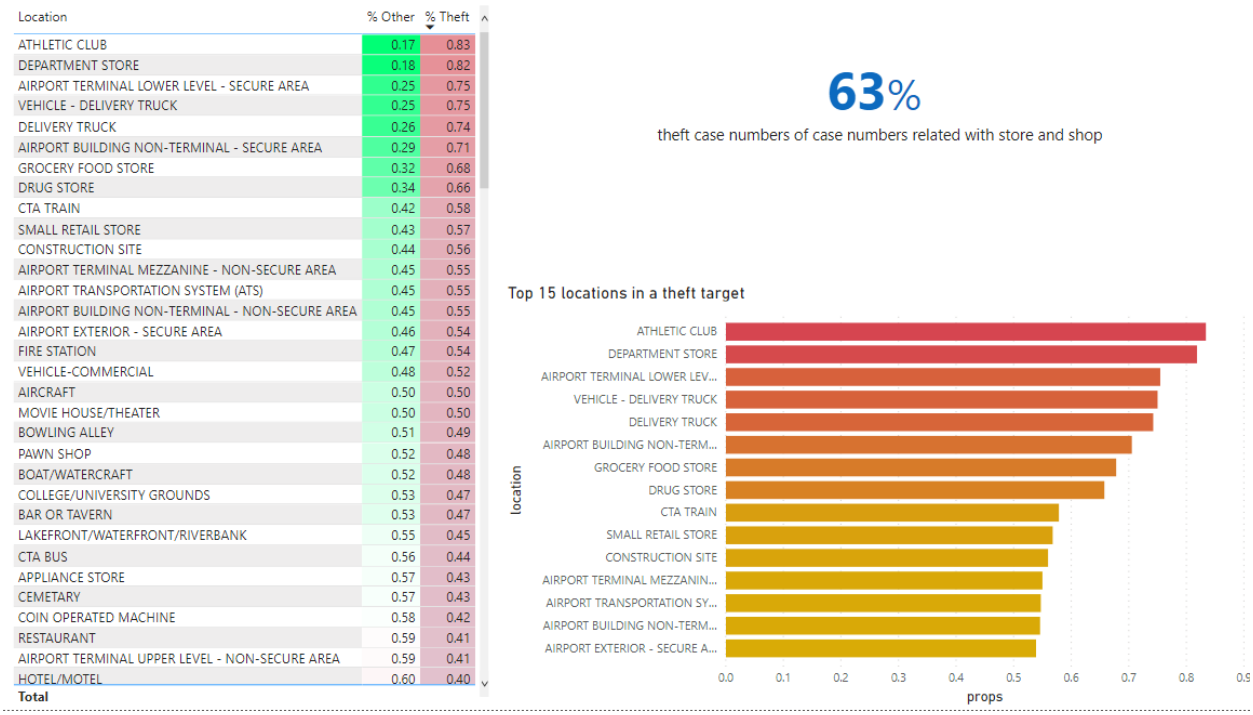

Figure 24 Analysis theft behaviors using location data.

This dashboard illustrates information about burglary location. It is useful to know which place can be targeted by theft crimes.

Results include:

- Almost the violent cases occurred at places which have a large of people and low security such as athletic club, department store, airport terminal lower level, …
- The burglary shows high interest in the stores and shops include department store, grocery food store, drug store, small retail store, … with 63% of total cases.

Location	% Other	% Theft
ATHLETIC CLUB	0.17	0.83
DEPARTMENT STORE	0.18	0.82
AIRPORT TERMINAL LOWER LEVEL - SECURE AREA	0.25	0.75
VEHICLE - DELIVERY TRUCK	0.25	0.75
DELIVERY TRUCK	0.26	0.74
AIRPORT BUILDING NON-TERMINAL - SECURE AREA	0.29	0.71
GROCERY FOOD STORE	0.32	0.68
DRUG STORE	0.34	0.66
CTA TRAIN	0.42	0.58
SMALL RETAIL STORE	0.43	0.57
CONSTRUCTION SITE	0.44	0.56
AIRPORT TERMINAL MEZZANINE - NON-SECURE AREA	0.45	0.55
AIRPORT TRANSPORTATION SYSTEM (ATS)	0.45	0.55
AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA	0.45	0.55
AIRPORT EXTERIOR - SECURE AREA	0.46	0.54
FIRE STATION	0.47	0.54
VEHICLE-COMMERCIAL	0.48	0.52
AIRCRAFT	0.50	0.50
MOVIE HOUSE/THEATER	0.50	0.50
BOWLING ALLEY	0.51	0.49
PAWN SHOP	0.52	0.48

Figure 25 Location names have a high percent of theft case numbers.

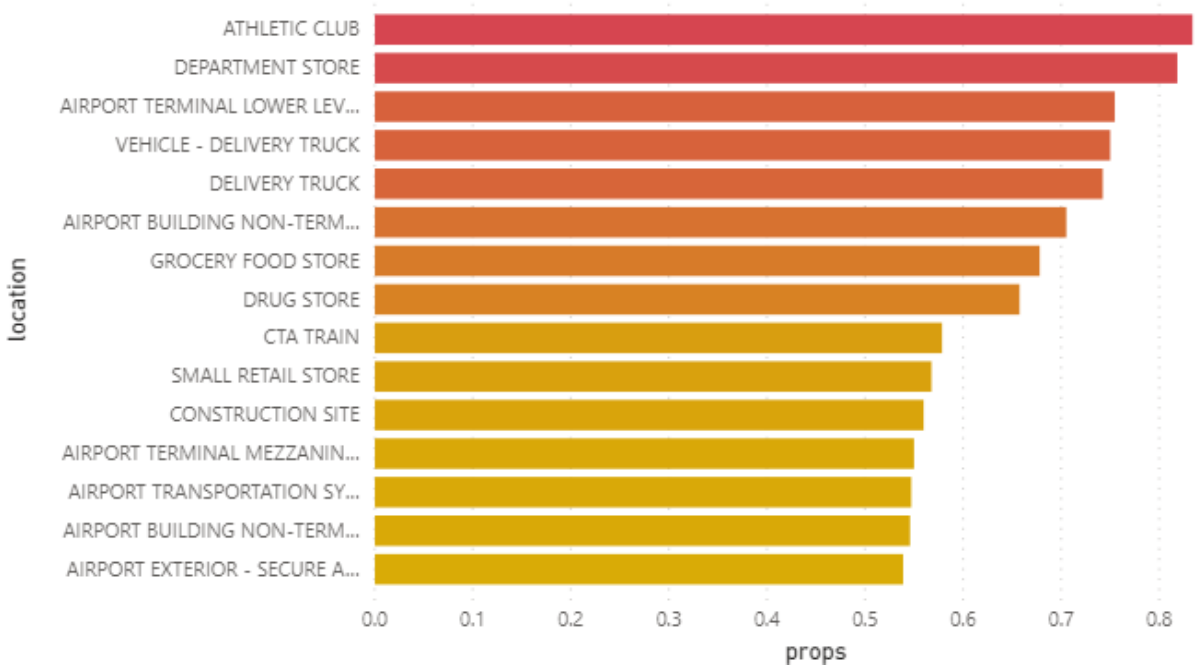

Figure 26 Top 15 location names in a theft target

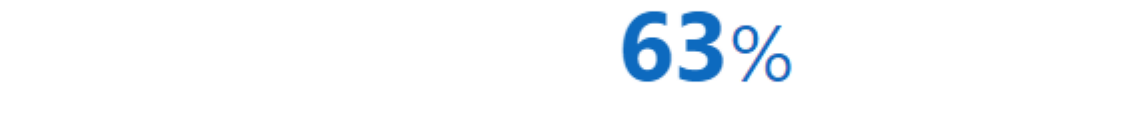

Figure 27 Store and shop which have percent of burglary

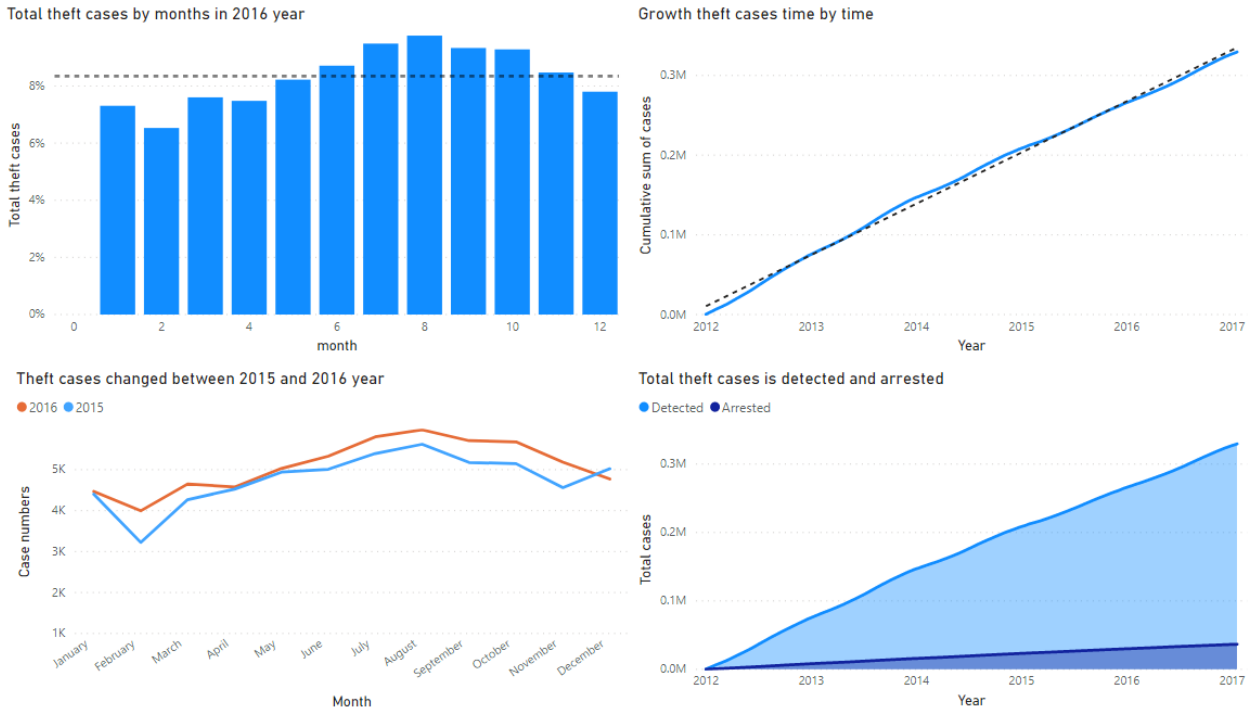

Figure 28 Analysis theft behaviors using time-series data.

This dashboard shows that the burglary activities by time-series to understand trends and occurrences. The burglary activities increased slightly in June and hit a peak in August with around 8.6% (Figure 29). The trend is predicted on the assumption that the theft crime rate decreased at the first months of the year and rise at June, July month. This was evidenced by comparing between 2015 and 2016 year (Figure 30).

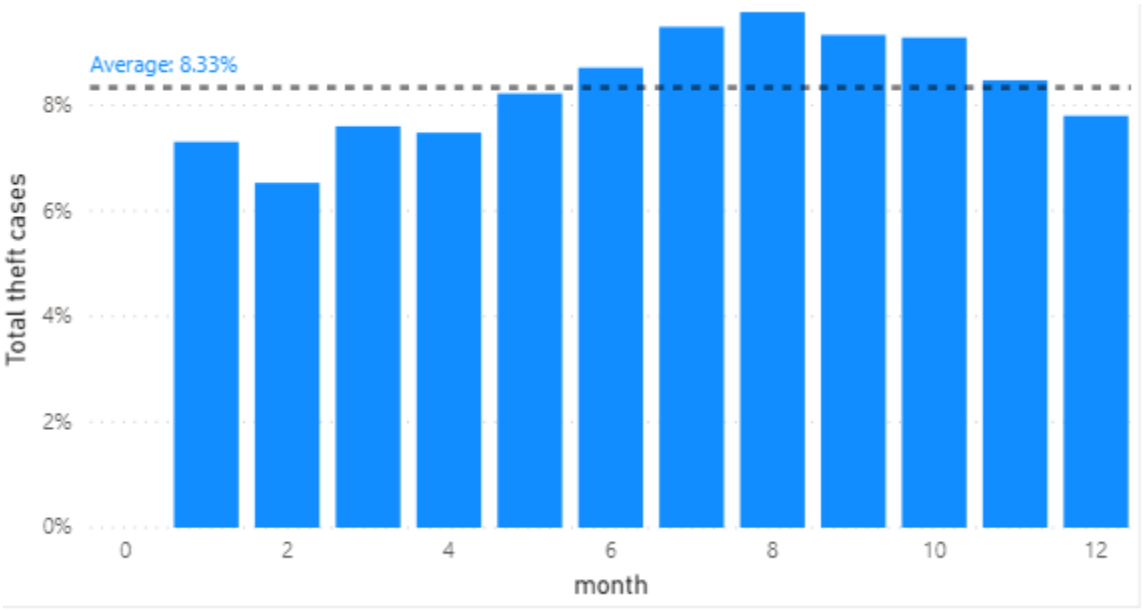

Figure 29 Total theft case numbers over a period of 12 months in 2016

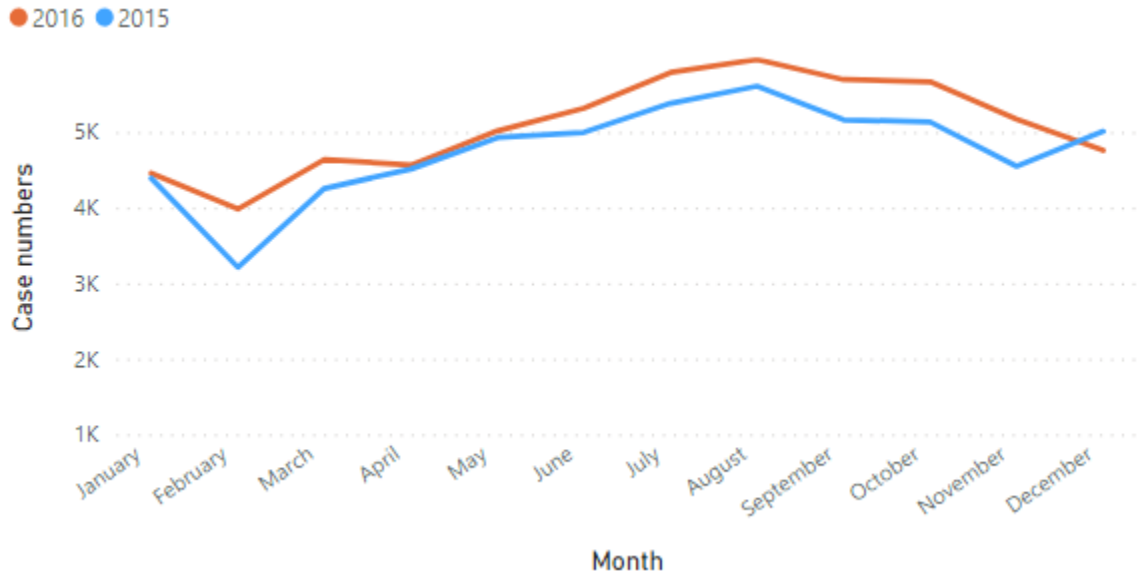

Figure 30 Theft cases changed between 2015 and 2016 according to data

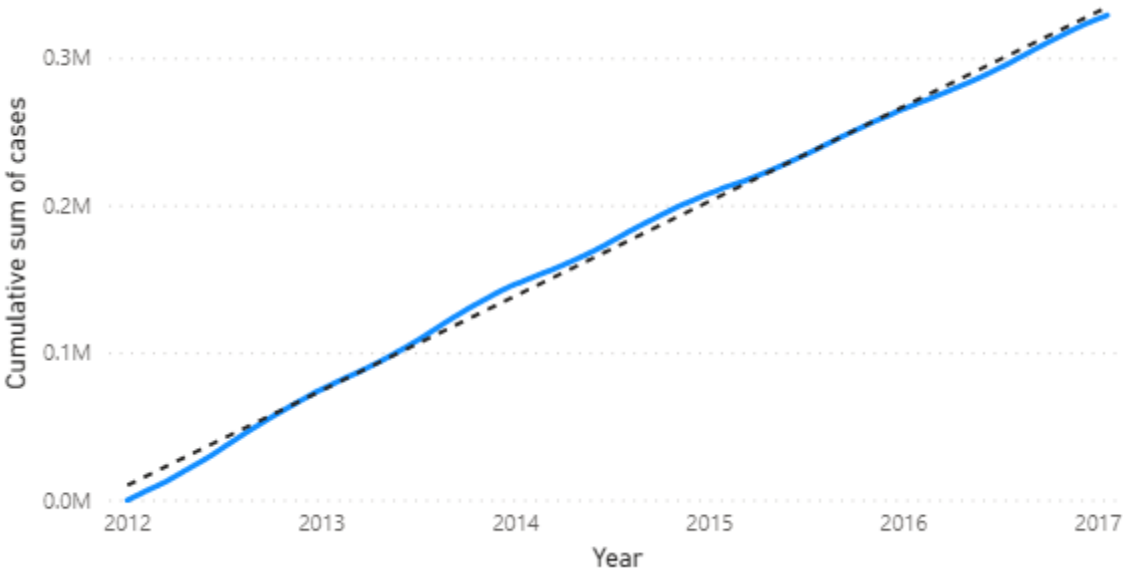

Figure 31 Growth theft case numbers over the past years

The figure 31 illustrates cumulative increasing of total cases by time. It almost covers linear regression line, perhaps theft crimes rate can be forecasted by using linear methods.

While reported burglary cases surge significant over past years (Figure 32), the number of arrested cases only marked approximately 30K in year 2017. It is explained that almost burglary cases were not critical incidents, so the policies spend more time to crime types such as murder, bank robbery, …

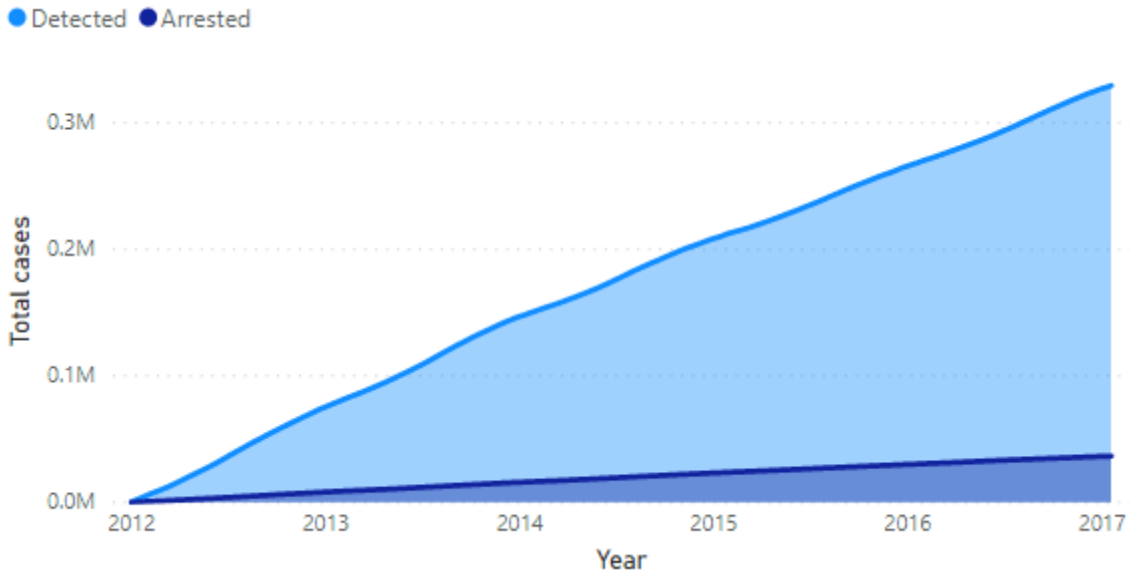

Figure 32 Total theft case numbers is detected and arrested

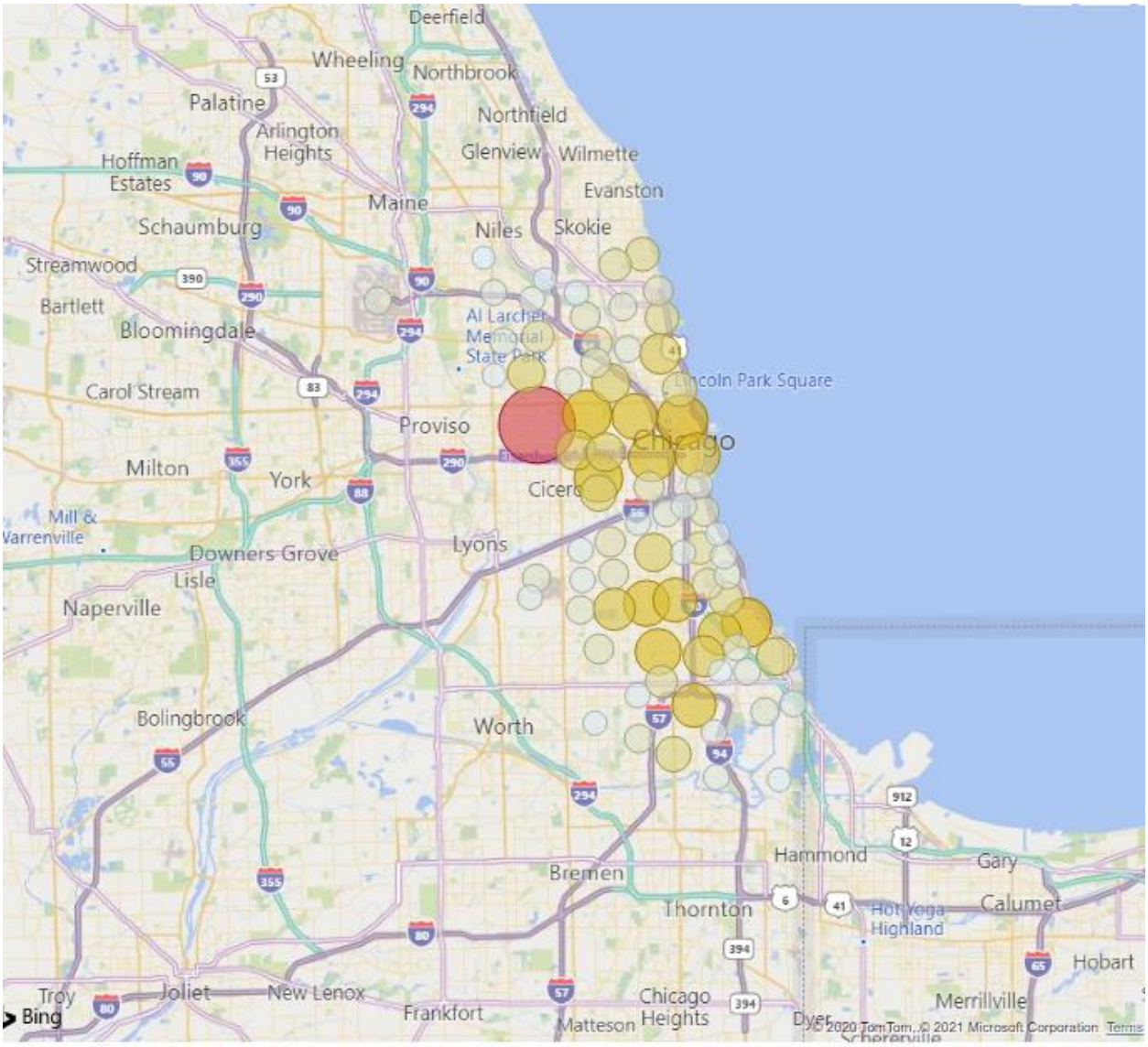

Figure 33 Number of theft cases visualize on map.

Briefly, there was 2 groups with large of cities gathering around. An idea for this visualization that using clustering algorithms to group cities to classify and analysis more detail. This report ignores that proposal because it does not associate with business objectives.

OLAP

To can go up and drill down when view data, a cube should be built with dimension hierarchies.

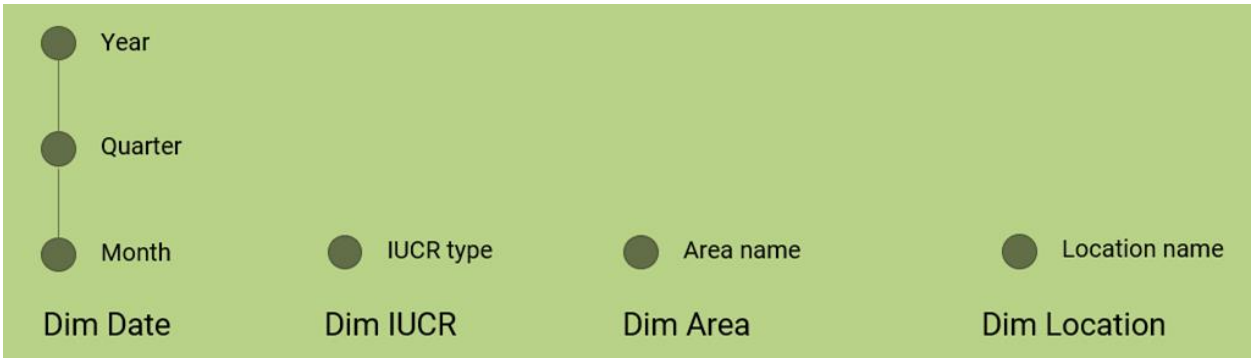

Figure 34 Dimension hierarchy on cube

There are 2 methods to retrieve measures from a cube:

- Use browser tab on SSAS
- Use MDX to query from cube

SSAS project load data from DDS to build a cube without changing any properties. The diagram be like below image.

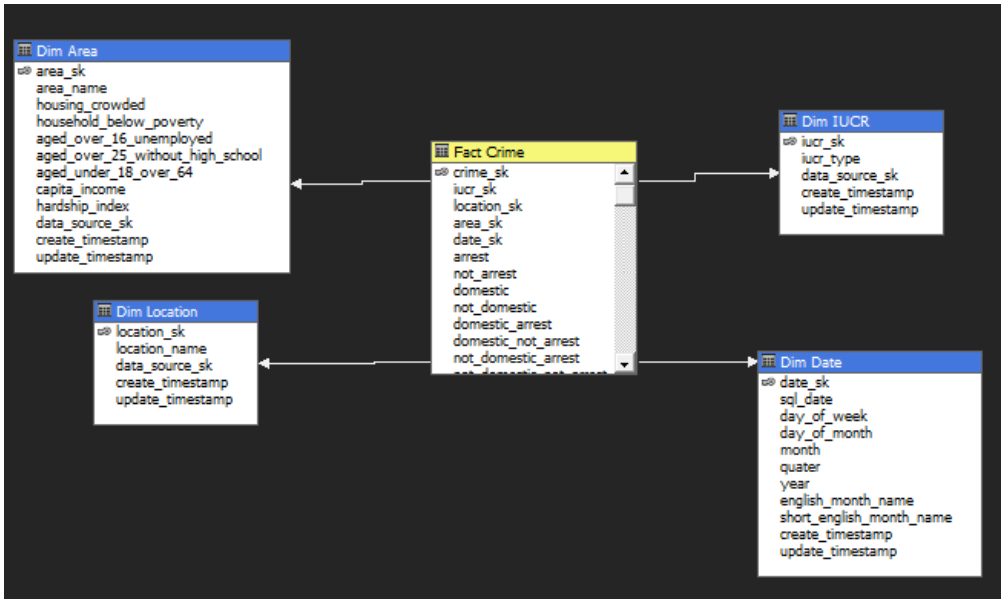

Figure 35 OLAP cube design

```
SELECT
NON EMPTY {[Measures].[Total cases],
           [Measures].[Arrest],
           [Measures].[Domestic]} ON COLUMNS,
NON EMPTY {
    [Dim Date].[Hierarchy].YEAR,
    [Dim Date].[Hierarchy]
} ON ROWS
FROM [Chicago Crimes DDS];
```

Figure 36 MDX query to calculate total cases, arrest and domestic

	Total cases	Arrest	Domestic
2012	75454	8244	2212
2013	71524	7726	2043
2014	61530	7356	1892
2015	57292	6727	2061
2016	61167	6387	2237
2017	2493	233	74
All	329460	36673	10519

Figure 37 Result retrieving

```
SELECT
NON EMPTY {[Dim Date].[Hierarchy].YEAR,
           [Dim Date].[Hierarchy]} ON COLUMNS,
NON EMPTY {[Dim IUCR].[Iucr Type].[Iucr Type]} ON ROWS
FROM [Chicago Crimes DDS]
WHERE [Measures].[Total cases];
```

Figure 38 MDX query to calculate total cases by year for each IUCR classification

	2012	2013	2014	2015	2016	2017	All
OTHER	260216	235179	212997	205703	204295	8864	1127254
THEFT	75454	71524	61530	57292	61167	2493	329460

Figure 39 Result retrieving

```
WITH MEMBER
    [Measures].[International] AS
    [Measures].[Not Domestic]
SELECT
NON EMPTY{
        [Measures].[Total cases],
        [Measures].[Arrest],
        [Measures].[International]
} ON COLUMNS,
NON EMPTY {
    ORDER(
        TOPCOUNT(
            [Dim Area].[Area Name].ALLMEMBERS, 10
        ),
        [Measures].[Total cases],
        DESC
    )
} ON ROWS
FROM [Chicago Crimes DDS]
;
```

Figure 40 MDX query to get top 10 cities saw highest theft crime rate

	Total cases	Arrest	International
All	1456714	377472	1236660
Austin	94730	35356	76890
Auburn Gresham	41634	11147	32720
Belmont Cragin	26791	6468	22371
Ashburn	13311	2215	11478
Albany Park	13040	2479	11192
Avondale	12936	2245	11181
Avalon Park	7667	1542	6345
Armour Square	6190	1334	5634
Archer Heights	5356	1039	4781

Figure 41 Result retrieving

Data mining

Extract hidden pattern from data is important aim at data mining. It can explain many questions which report and OLAP cannot. It provides a insight from natural properties datasets.

Which factors affected to theft behaviors?

To determine area factors, include income, population, age, and employee status whether affected to theft behaviors or not, there 2 approaches:

- Analysis by use PCA method
- Analysis by use linear regression

PCA visualization

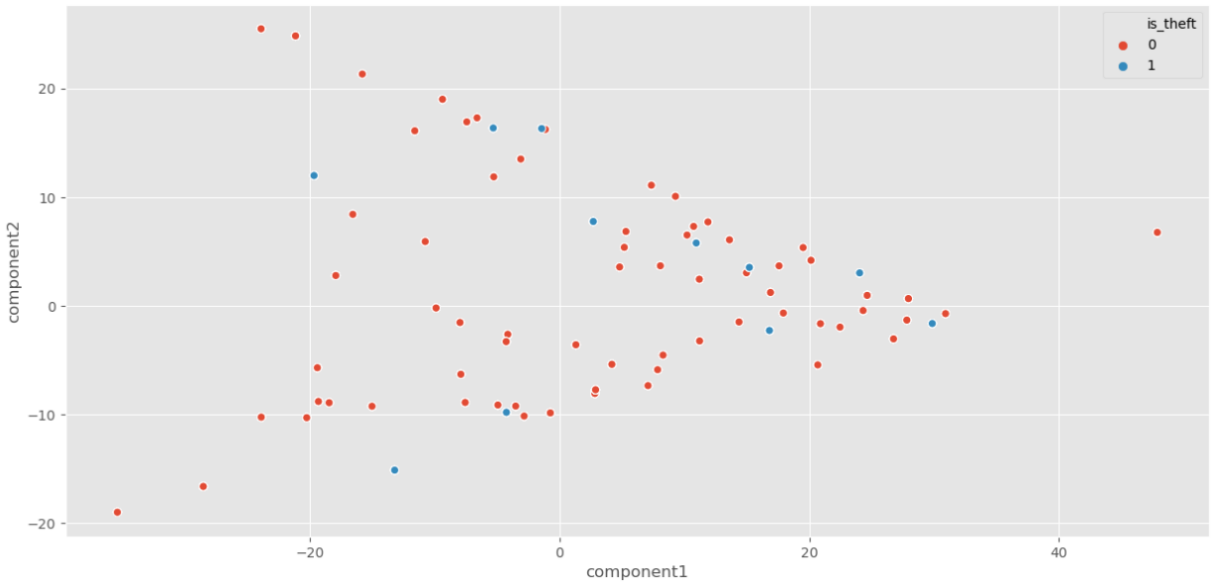

Figure 42 Visualize community area features by reduce dimension numbers to 2

PCA select 2 dimensions have highest of variances from a set of features. However, only 2 dimensions cannot classify theft category by linear or non-linear models. So, to extract hidden pattern from community area features, linear regression methodology should be applied.

Linear regression

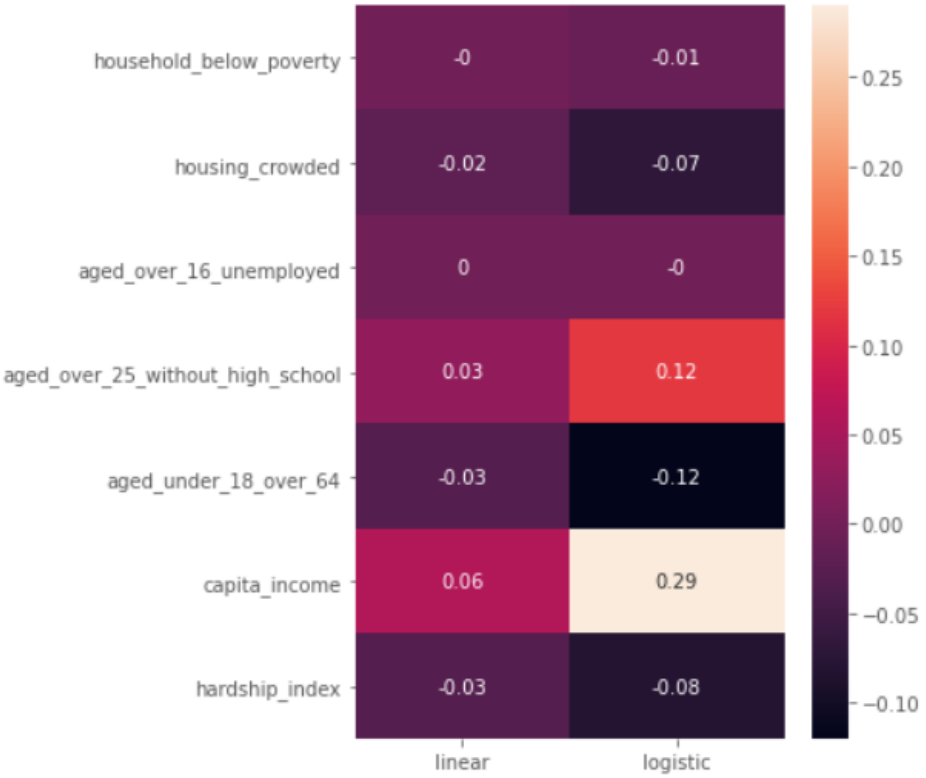

Figure 43 Coefficient each variable by linear and logistic regression

Linear and logistic regression is 2 most popular models is used to analysis data with many variables. This report represents both approaches to get more confidence in conclusion. Depend on coefficient of 7 variables, logistic regression model has a better significant than linear regression, so it is used to explain variables meaning at the end.

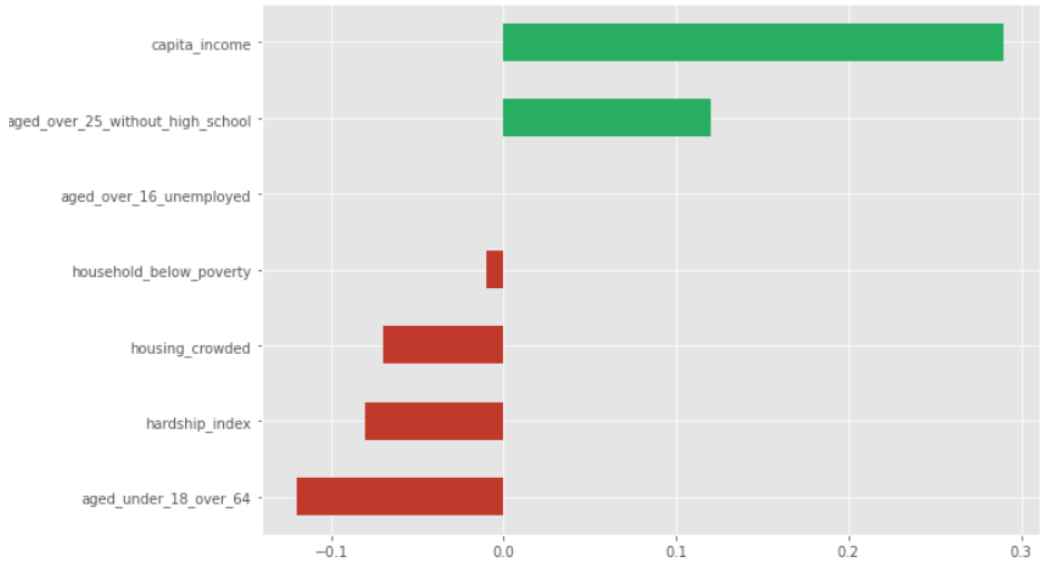

Figure 44 Coefficient of variables by logistic regression

The final function of are features which affect to theft behaviors.

$$0.29 * X1 + 0.12 * X2 - 0.07 * X3 - 0.08 * X4 - 0.12 * X5$$

X1: capita income
X2: aged over 25 without high school
X3: housing crowded
X4: hardship index
X5: aged under 18 over 64

By the formula, capita income and percentage of aged over 25 without high school **positive correlation** with theft category. By contrast, percentage of housing crowded, percentage of aged under 18 or over 64 and hardship index **negative correlation** relationship with theft crimes.

Predict a case is whether theft category or not?

Find out which important features is a key step to build a model. The first model is built from 16 features which are collected from DDS. It is called baseline model with random forest strategy. The second model use a feature which present a case happened whether at the shop or not. Also, it uses interaction variables from existed columns. Finally, a best model is chosen between Random forest and KNN.

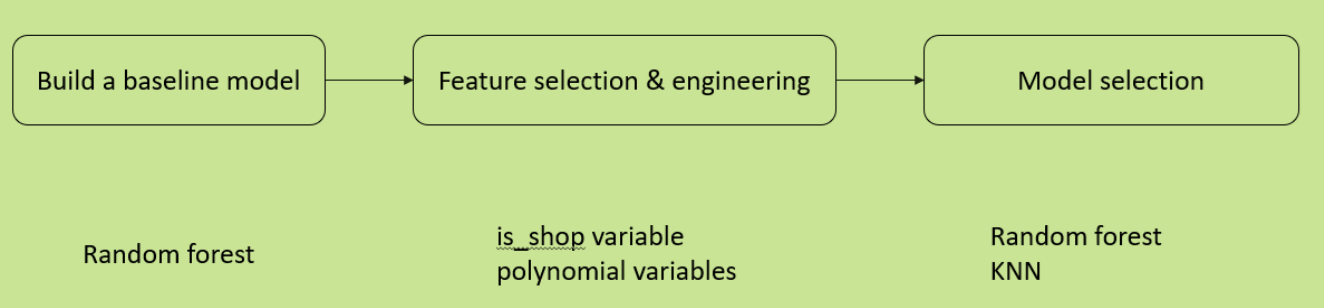

Figure 45 Data mining processes

Model	recall	precision	f1 score
Random forest with 16 existed features	0.59	0.14	0.23
Random forest with 16 existed features + is_shop	0.64	0.22	0.33
Random forest with 16 existed features + is_shop + interaction features	0.62	0.25	0.36
KNN with 16 existed features + is_shop + interaction features	0.55	0.22	0.31

Figure 46 Model evaluation

Random forest model has better score than KNN with marking 0.33 f1. The final model is third model after optimizing parameters.

APPENDIX A – Datasets from sources

Chicago Crimes dataset (2012 – 2017)

#	Feature	Description	Sample value
1	ID	The unique identifier for a record	10508693
2	Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	HZ250496
3	Date	Date when the incident occurred. this is sometimes a best estimate.	2016-05-03 23:40:00.000
4	Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.	013XX S SAWYER AVE
5	IUCR	The Illinois Uniform Crime Reporting code	0486
6	Primary Type	The primary description of the IUCR code	BATTERY
7	Description	The secondary description of the IUCR code, a subcategory of the primary description.	DOMESTIC BATTERY SIMPLE
8	Location Description	Description of the location where the incident occurred.	APARTMENT
9	Arrest	Indicates whether an arrest was made.	1
10	Domestic	Indicates whether the incident was domestic related as defined by the Illinois Domestic Violence Act.	1
11	Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.	1022
12	District	Indicates the police district where the incident occurred.	10
13	Ward	The ward (City Council district) where the incident occurred.	24
14	Community Area	Indicates the community area where the incident occurred.	29
15	FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).	08B
16	X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.	1154907
17	Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.	1893681
18	Year	Year the incident occurred.	2016

19	Updated On	Date and time the record was last updated.	2016-05-10 15:56:50.000
20	Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	41.86407
21	Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	-87.70682
22	Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.	(41.864073157, -87.706818608)

FBI Code

	Column	Description	Sample value
1	Code	FBI code unique	01A
2	Name	A name is used to classifier for crime	Larceny
3	Type	Which category of name	Crimes against property
4	Def	Name definition	Definition: The unlawful taking, carrying, leading, or riding away of property from the possession or constructive possession of another person.

Socioeconomic

#	Column	Description	Sample value
1	Community Area Number	A unique identifier	24
2	Community area name	A community name	Wes Town
3	Percent of housing crowded	Percent occupied housing units more than one person per room	3.8
4	Percent of households below poverty	Percent of households living below the federal poverty level	24
5	Percent aged 16+ unemployed	Percent of persons over the age of 16 years that are unemployed	7
6	Percent aged 25+ without high school diploma	Percent of persons over the age of 25 years without a high education	13.4
7	Percent aged under 18 or over 64	Percent of the population under 18 or over 64 years of age	27.5
8	Per capita income	Community area per capita income is estimated as the sum of tract-level aggregate incomes divided by the total population	23939
9	Hardship index	Score that incorporates each of the six selected socioeconomic indicators	1

APPENDIX B – Transformation from stage to NDS

Data source

Variable	Description	Source	Transformation
data_source_sk	A surrogate key	Chicago Crimes FBI Code Socioeconomic	Union operator from sources
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

FBI Code NDS

Variable	Description	Source	Transformation
fbi_type_sk	A surrogate key	___	___
fbi_type_name	Type name of FBI code	FBI Code	Get unique values from type variable
data_source_sk	Reference to data source	FBI Code	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

FBI NDS

Variable	Description	Source	Transformation
fbi_code_sk	A surrogate key	___	___
fbi_code_nk	Natural key of FBI code	FBI Code	Get values from fbi_code variable
fbi_code_name	FBI code name	FBI Code	Get values from name variable
fbi_code_type	Reference to FBI Type NDS	FBI Code FBI Type NDS	Join with FBI Type NDS by type and get values from fbi_code_sk
fbi_type_definition	Definition for each FBI code	FBI code	Gete values from Def variable
data_source_sk	Reference to data source	FBI Code	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Primary category NDS

Variable	Description	Source	Transformation
primary_category_sk	A surrogate key	___	___
primary_category_name	Category name of IUCR code	Chicago Crimes	Get unique values from primary type variable
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Secondary category NDS

Variable	Description	Source	Transformation
secondary_category_sk	A surrogate key	___	___
secondary_category_name	Secondary category name of IUCR code	Chicago Crimes	Get unique values from description variable
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

IUCR NDS

Variable	Description	Source	Transformation
iucr_sk	A surrogate key	___	___
iucr_nk	Natural key of IUCR	Chicago Crimes	Get unique values from IUCR variable
primary_category_sk	Category key of IUCR code	Chicago Crimes Primary category NDS	Join [Chicago Crimes] with [Primary category NDS] and derived from primary_category_sk
secondary_category_sk	Secondary category key of IUCR code	Chicago Crimes Secondary category NDS	Join [Chicago Crimes] with [Secondary category NDS] and derived from secondary_category_sk
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Area NDS

Variable	Description	Source	Transformation
area_sk	A surrogate key	___	___
area_nk	Natural key of Community area	Socioeconomic	Get values from Community area number
area_name	Community area name	Socioeconomic	Get values from community area name
housing_crowded	Percent occupied housing units more than one person per room	Socioeconomic	Get values from Percent of housing crowded
household_below_poverty	Percent of households living below the federal poverty level	Socioeconomic	Get values from Percent of household below poverty
aged_over_16_unemployed	Percent of persons over the age of 16 years that are unemployed	Socioeconomic	Get values from Percent of 16+ unemployed
aged_over_25_without_highschool	Percent of persons over the age of 25 years without a high education	Socioeconomic	Get values from Percent aged 25+ without high school diploma
aged_under_18_over_64	Percent of the population under 18 or over 64 years of age	Socioeconomic	Get values from percent aged under 18 or over 64
capita_income	Community area per capita income is	Socioeconomic	Get values from Per capita income

	estimated as the sum of tract-level aggregate incomes divided by the total population		
hardship_index	Score that incorporates each of the six selected socioeconomic indicators	Socioeconomic	Get values from Hardship index
data_source_sk	Reference to data source	Socioeconomic	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Address NDS

Variable	Description	Source	Transformation
address_sk	A surrogate key	___	___
address_name	Address name where a crime is detected	Chicago Crimes	Get unique values from Block variable
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Location NDS

Variable	Description	Source	Transformation
location_sk	A surrogate key	___	___
location_name	Location name where a crime is detected	Chicago Crimes	Get unique values from location description variable
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

District NDS

Variable	Description	Source	Transformation
district_sk	A surrogate key	___	___
district_nk	An actual value of district	Chicago Crimes	Get unique values from district variable
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Beat NDS

Variable	Description	Source	Transformation
beat_sk	A surrogate key	___	___
beat_nk	An actual value of beat	Chicago Crimes	Get unique values from beat variable
district_sk	Reference to District NDS	Chicago Crimes District NDS	Join [Chicago Crimes] with [District NDS] on district variable and get values from district_sk
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Date NDS

Variable	Description	Source	Transformation
date_sk	A surrogate key	___	___
date_nk	A value is stored in SQL	Chicago Crimes	Get unique values from date
day	Day name	Chicago Crimes	Get values from DAY(date_nk)
month	Month number	Chicago Crimes	Get values from MONTH(date_nk)
year	Year number	Chicago Crimes	Get values from YEAR(date_nk)
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Crimes NDS

Variable	Description	Source	Transformation
crimes_sk	A surrogate key	___	___
crimes_nk	Natural key	Chicago Crimes	Get values from ID variable
case_number	Number case of crimes	Chicago Crimes	Get values from Case number
iuck_sk	Reference to IUCR NDS	Chicago Crimes IUCR NDS	Join [Chicago Crimes] with [IUCR NDS] on iucr variable and get values from iucr_sk
address_name	Address name where a crime is detected	Chicago Crimes	Get unique values from Block variable
data_source_sk	Reference to data source	Chicago Crimes	Get from source_id
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

APPENDIX C – Transform from NDS to DDS

Dim Date

Variable	Description	Sample value	Transformation
date_sk	Surrogate key	1	Get date_sk from Date NDS
sql_date	Date value is stored in database	1999-01-01 15:30:00	Get date_nk from Date NDS
day_of_week	Day order by week	2	Get day from Date NDS
day_of_month	Day order by month	20	DATENAME(dw, date_nk)
month	Month number	8	Get month from Date NDS
quarter	Quarter in year	3	DATEPART(q, date_nk)
year	Year	2012	Get year from Date NDS
english_month_name	Month name	August	DATENAME(month, date_nk)
short_english_month_name	Short month name	Aug	SUBSTRING(DATENAME(month, date_nk), 1, 3)
data_source_sk	Data source key from NDS	1	Get data_source_sk from Date NDS
create_timestamp	Represent created time in DDS	___	___
update_timestamp	Represent updated time in DDS	___	___

Dim Location

Variable	Description	Sample value	Transformation
location_sk	Surrogate key	1	Get location_sk from Location NDS
location_name	Name location of the crime	Hotel/Motel	Get location_name from Location NDS
data_source_sk	Data source key from NDS	1	Get data_source_sk from Date NDS
create_timestamp	Represent created time in DDS	___	___
update_timestamp	Represent updated time in DDS	___	___

Dim Area

Variable	Description	Sample value	Transformation
area_sk	A surrogate key	24	Get area_sk from Area NDS
area_name	Community area name	Wes Town	Get area name from Area NDS
housing_crowded	Percent occupied housing units more than one person per room	3.8	Get housing_crowded from Area NDS
household_below_poverty	Percent of households living below the federal poverty level	24	Get household_below_poverty from Area NDS
aged_over_16_unemployed	Percent of persons over the age of 16 years that are unemployed	7	Get aged_over_16_unemployed from Area NDS
aged_over_25_without_highschool	Percent of persons over the age of 25 years without a high education	13.4	Get aged_over_25_without_high school from Area NDS
aged_under_18_over_64	Percent of the population under 18 or over 64 years of age	27.5	Get aged_under_18_year_over_64 from Area NDS
capita_income	Community area per capita income is estimated as the sum of tract-level aggregate incomes divided by the total population	23939	Get capita_income form Area NDS
hardship_index	Score that incorporates each of the six selected socioeconomic indicators	1	Get hardship_index form Area NDS
data_source_sk	Reference to data source	1	Get data_source_sk from Area NDS
create_timestamp	Represent created time in NDS	___	___
update_timestamp	Represent created time in NDS	___	___

Dim IUCR

Variable	Description	Sample value	Transformation
iucr_sk	A surrogate key	1	Get iucr_sk from IUCR NDS
is_theft	Is a theft or not	1	CASE WHEN iucr_type = 'THEFT' THEN 1 ELSE 0 END;
data_source_sk	Reference to data source	1	Get data_source_sk from IUCR NDS
create_timestamp	Represent created time in DDS	___	___
update_timestamp	Represent updated time in DDS	___	___

Fact Crime

Variable	Description	Sample value	Transformation
crimes_sk	A surrogate key	1	Get crimes_sk from Crime NDS
date_sk	Surrogate key	1	Get date_sk from Crime NDS
location_sk	Surrogate key	1	Get location_sk from Crime NDS
iuck_sk	Reference to IUCR NDS	1	Get iucr_sk from Crime NDS
area_sk	A surrogate key	1	Get area_sk from Crime NDS
arrest	Have been arrested	1	Get arrest from Crime NDS
not_arrest	Haven't been arrested	0	1 – arrest
domestic	Domestic criminal	1	Get domestic from Crime NDS
not_domestic	Not domestic criminal	0	1 – domestic
domestic_arrest	Domestic criminal have been arrested	1	CASE WHEN arrest = 1 AND domestic = 1 THEN 1 ELSE 0 END;

domestic_not_arrest	Domestic criminal haven't been arrested	0	CASE WHEN arrest = 0 AND domestic = 1 THEN 1 ELSE 0 END;
not_domestic_not_arrest	Not Domestic criminal haven't been arrested	0	CASE WHEN arrest = 0 AND domestic = 0 THEN 1 ELSE 0 END;
not_domestic_arrest	Not Domestic criminal have been arrested	0	CASE WHEN arrest = 1 AND domestic = 0 THEN 1 ELSE 0 END;
data_source_sk	Reference to data source	1	Get data_source_sk from Crime NDS
create_timestamp	Represent created time in DDS	___	___
update_timestamp	Represent updated time in DDS	___	___

Variables in SSIS

Variable	Data Type	Description	Sample Value
ExecutionTime	DateTime	Get current time for each time executing data. Get value from current_time_execution at CurrentTime Table	1/1/1999
NDSLastUpdated	DateTime	Get the latest update time of NDS. Get value from the biggest updated timestamp at LastUpdated Table with process's name 'NDS weekly incremental'	1/1/1999
DDSLastUpdated	DateTime	Get the latest update time of DDS. Get value from the biggest updated timestamp at LastUpdated Table with process's name 'DDS monthly incremental'	1/1/1999

APPENDIX D – Best practices on SSIS

1. How to join 2 tables in SSIS?

A lookup component should be used to join 2 tables together. To improve performance for matching, SSIS support 3 mode:

- Full cache: The system automatically cache result from query to reuse in the next time before the packages executed completed.
- Partial cache: Only cache primary key to support query.
- No cache: Same as its name, this mode makes no data is cached.

Almost in the whole situation, a full cache model should be used to improve performances packages. However, full cache mode compares 2 strings with uppercase and space difference instead of ignoring them as SQL Server. When use this mode, to ensure data flow is executed right, you must uppercase and trim all spaces in strings comparing.

2. How to reduce time when populate data from NDS to DDS in the first time?

A large of data will move from NDS to DDS in the first time. If you use SCD component, it takes the long time to finish executing. The trick can be used in here that you can disable SCD because the destination DDS no data in the first time, you do not need check existed or update anything. You can enable them after first time running.