Project 1 – CITS3403: Data Warehousing

Association Rule Mining

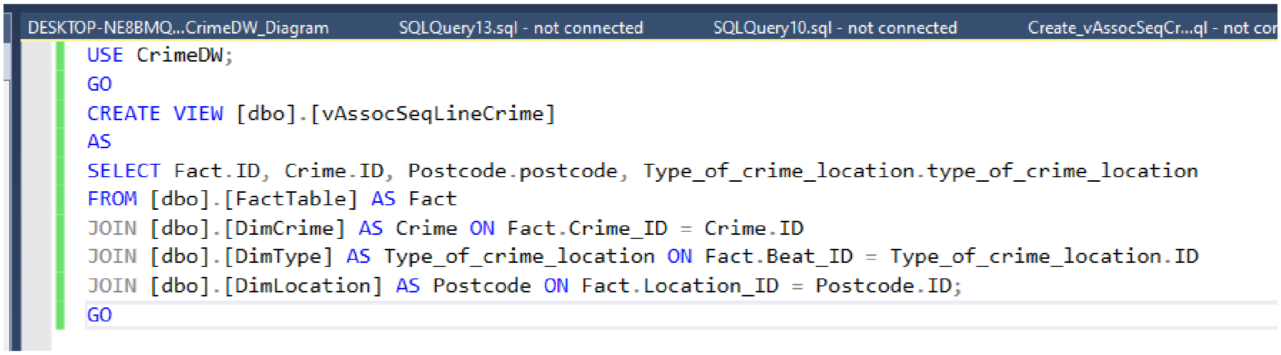
Henry Tran - 23035141

A. Associate rule mining process:

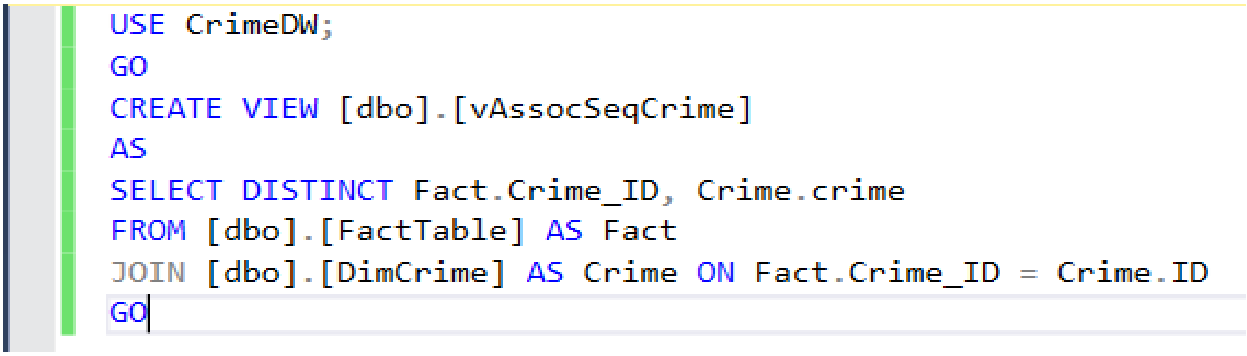
In this paper, I will show how I did for the association rule mining for this crime dataset

The software I used in this association rule mining is Visual Studio 2019, SQL Server Management Studio (SSMS) and SQL Server Analysis Service Engine.

1. Create the views in SSMS:
   - As I got my data stored in SSMS, I thought about the rule mining we could find out from this data set. Therefore, I found 2 objectives of this association rule mining, they are:
     - Which type of crime location can a crime happen together
     - If a crime happened in postcode A and/or postcode B, that crime can also happen in postcode C.
   - From my objectives, I must have 2 tables, one table would contain a unique crime key and the name of the crime (I call it Crime) and another table would contain a single record of crimes with a unique key along with crime foreign key and information about postcode and type of location that a crime happened (I call it LineCrime).
   - After that, I created 2 views on SSMS, here is the screenshots of my scripts. All 2 views will have the same table "start": vAssocSeq along with LineCrime/Crime

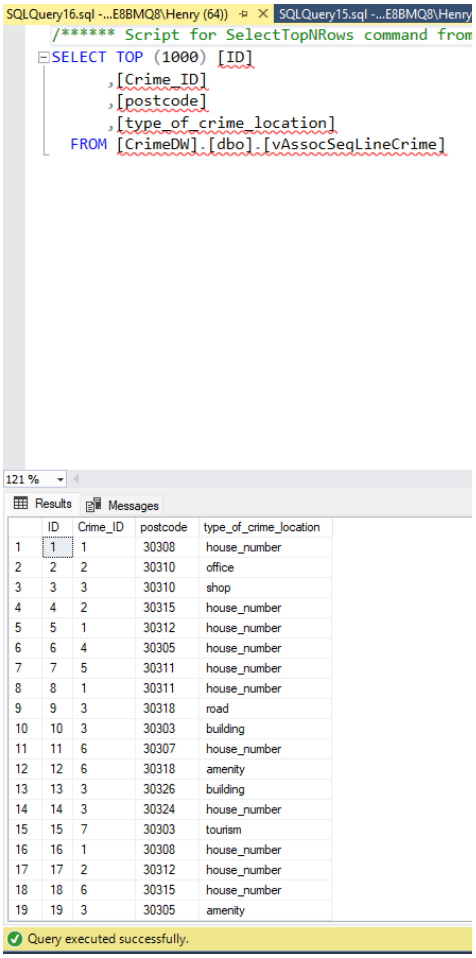

DESKTOP-NE8BMQ...CrimeDW_Diagram
SQLQuery13.sql - not connected
SQLQuery10.sql - not connected
Create_vAssocSeqCr...ql - not cor

```sql
USE CrimeDW;
GO
CREATE VIEW [dbo].[vAssocSeqLineCrime]
AS
SELECT Fact.ID, Crime.ID, Postcode.postcode, Type_of_crime_location.type_of_crime_location
FROM [dbo].[FactTable] AS Fact
JOIN [dbo].[DimCrime] AS Crime ON Fact.Crime_ID = Crime.ID
JOIN [dbo].[DimType] AS Type_of_crime_location ON Fact.Beat_ID = Type_of_crime_location.ID
JOIN [dbo].[DimLocation] AS Postcode ON Fact.Location_ID = Postcode.ID;
GO
```



```
USE CrimeDW;
GO
CREATE VIEW [dbo].[vAssocSeqCrime]
AS
SELECT DISTINCT Fact.Crime_ID, Crime.crime
FROM [dbo].[FactTable] AS Fact
JOIN [dbo].[DimCrime] AS Crime ON Fact.Crime_ID = Crime.ID
GO
```

Here is our result after we created the LineCrime view:



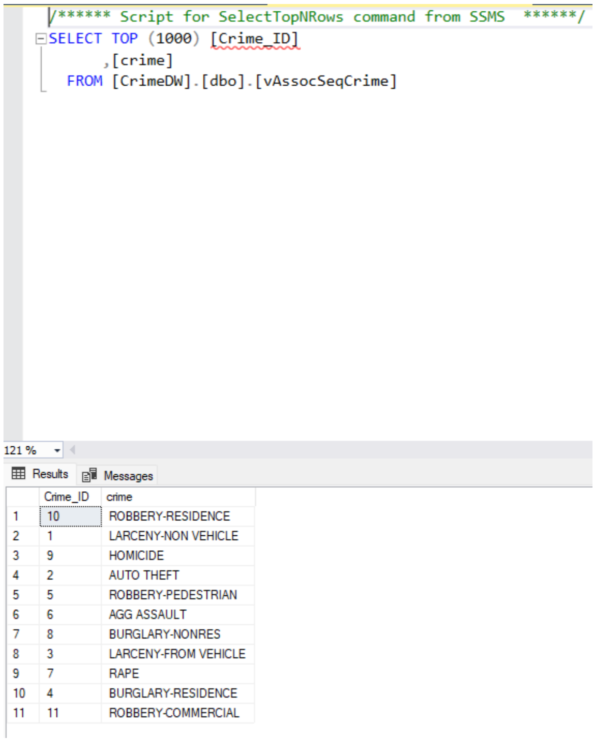
SQLQuery16.sql -...E8BMQ8\Henry (64))
SQLQuery15.sql -...E8BMQ8\Henry
121 %
Results
Messages

```
/****** Script for SelectTopNRows command from
SELECT TOP (1000) [ID]
      ,[Crime_ID]
      ,[postcode]
      ,[type_of_crime_location]
  FROM [CrimeDW].[dbo].[vAssocSeqLineCrime]
```

	ID	Crime_ID	postcode	type_of_crime_location
1	1	1	30308	house_number
2	2	2	30310	office
3	3	3	30310	shop
4	4	2	30315	house_number
5	5	1	30312	house_number
6	6	4	30305	house_number
7	7	5	30311	house_number
8	8	1	30311	house_number
9	9	3	30318	road
10	10	3	30303	building
11	11	6	30307	house_number
12	12	6	30318	amenity
13	13	3	30326	building
14	14	3	30324	house_number
15	15	7	30303	tourism
16	16	1	30308	house_number
17	17	2	30312	house_number
18	18	6	30315	house_number
19	19	3	30305	amenity

✓ Query executed successfully.

And here is the result after we created the Crime view:



```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [Crime_ID]
      ,[crime]
  FROM [CrimeDW].[dbo].[vAssocSeqCrime]
```

	Crime_ID	crime
1	10	ROBBERY-RESIDENCE
2	1	LARCENY-NON VEHICLE
3	9	HOMICIDE
4	2	AUTO THEFT
5	5	ROBBERY-PEDESTRIAN
6	6	AGG ASSAULT
7	8	BURGLARY-NONRES
8	3	LARCENY-FROM VEHICLE
9	7	RAPE
10	4	BURGLARY-RESIDENCE
11	11	ROBBERY-COMMERCIAL

Note: Should not create the view directly from the dimension table. The reason is that in different cases, the crime ID in the dimension table may not appear in our LineCrime view. In this case it is okay but just to mention in other cases.

2. <u>Create an Analysis Service Multi-dimensional Data Mining Process:</u>
   - I opened Visual Studio and created a project using Analysis Service Multi-dimensional Data Mining Project
   - I created project name: “Project1_DW_2023”
3. <u>Create a new Data Source for the association rule mining project</u>
4. <u>Create Data Source View:</u>
   - When I created the data source view, I added our 2 views.
   - After that, I had to add a relationship between the ID of the Crime view and Crime_ID of LineCrime view.
5. <u>Perform Association Rule Mining:</u>
   - I brought up the Data Mining Wizard and created mining structure with the Microsoft Association Rules technique.
   - In our data mining, we need to select which view is the case table, which view is the nested table. Because the case table’s primary key will be the nested table’s foreign key, therefore, Crime view was the case table and LineCrime was the nested table. (Case table contains unique value while nested table contains multiple rows of the same value)

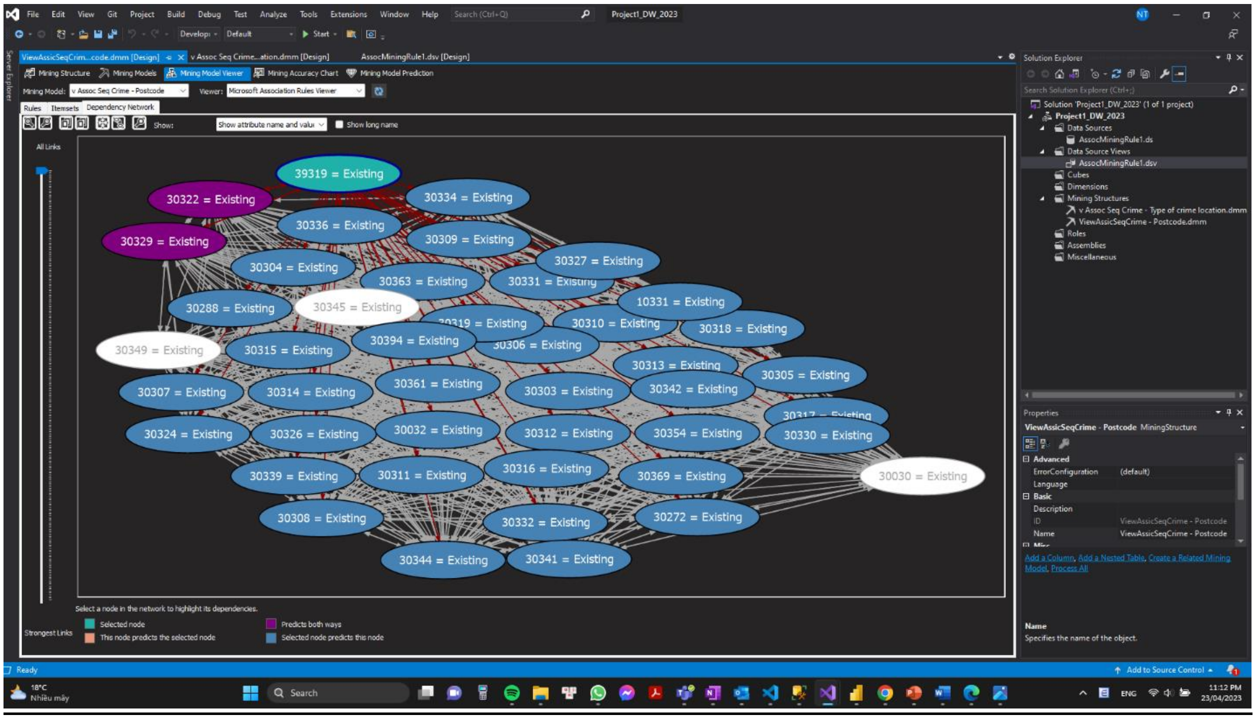
- We also need to find the single key column (represents each unique record – in this case, is the ID column from LineCrime view), a single predictable column (represent the value we will use to predict – in this case, we have 2 predictable columns: postcode and type_of_crime_location) and input column (the data contains in both 2 views – in this case, is the ID of the crime). We will create 2 data rule mining for each predictable column

6. <u>View the Association Rule Mining Rule</u>

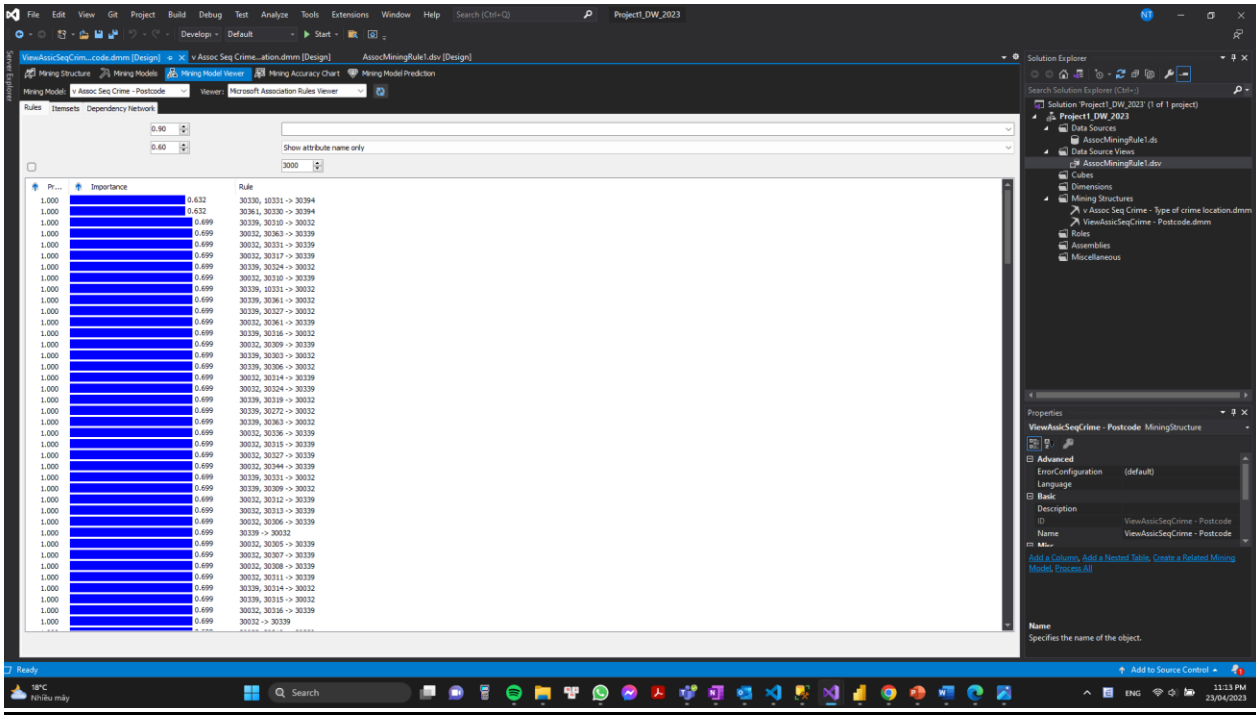
B. <u>Result:</u>

1. <u>Postcode data rule mining:</u>

- We got the dependency network. Here is an example of the postcode 39319. We can see that this postcode has a "2-way relationship" with postcodes 30329 and 30322, which means if a crime happened in either 1 or 2 postcodes, then the crime would happen in the remaining postcode among 3 of them. Most of the crime that happens in other postcodes could lead to a crime in postcode 39319. 3 postcodes 30349, 30345 and 30030 has no "relationship" with our selected postcode, which explains that if a crime happened in 1 or 2 postcodes among these 3 postcodes, then that crime would not happen in postcode 30319


Mining Model: v Assoc Seq Crime - Postcode
Viewer: Microsoft Association Rules Viewer
Show: Show attribute name and valu

- Looking at the rule mining model, we set the probability of 90% and importance of 60%, then we got many rules which satisfied the condition with 100%. For example, if a crime happened in postcodes 30330 and 10331, then the crime would happened in postcode 30394.

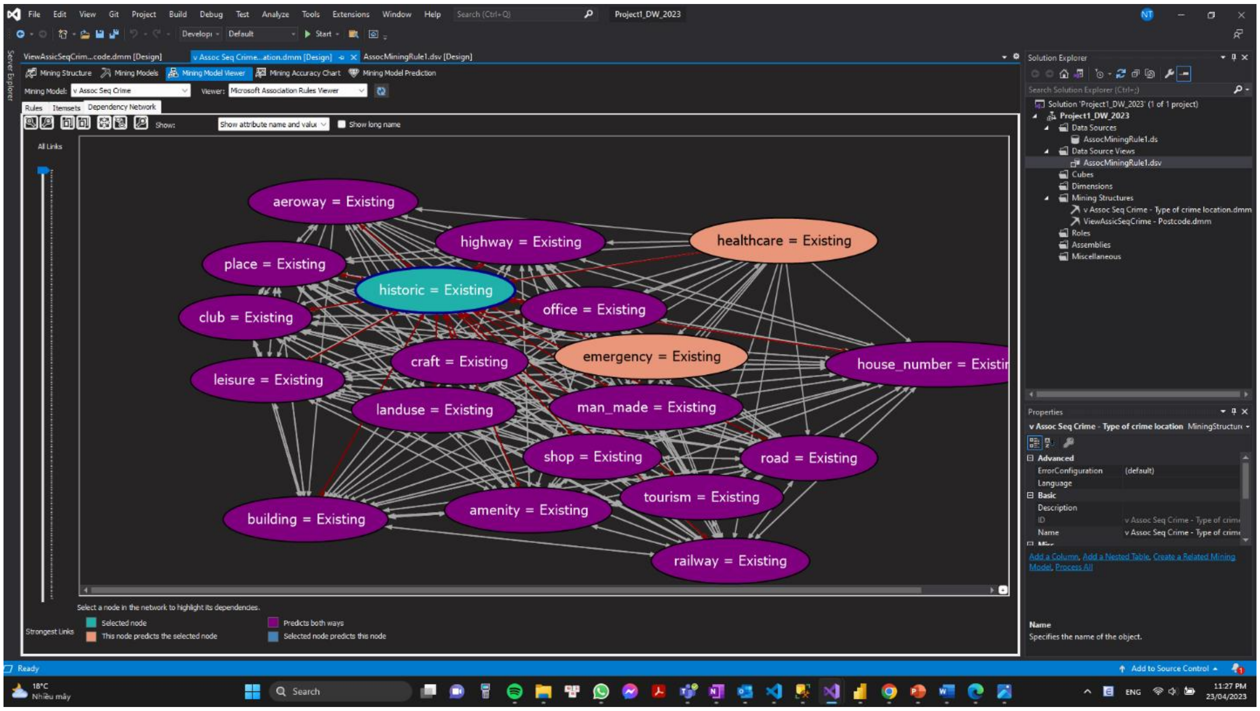


Mining Model: v Assoc Seq Crime - Postcode

Viewer: Microsoft Association Rules Viewer

Pr...	Importance	Rule
1.000	0.632	30330, 10331 -> 30394
1.000	0.632	30361, 30330 -> 30394
1.000	0.699	30339, 30310 -> 30032
1.000	0.699	30032, 30363 -> 30339
1.000	0.699	30032, 30331 -> 30339
1.000	0.699	30032, 30317 -> 30339
1.000	0.699	30339, 30324 -> 30032
1.000	0.699	30032, 30310 -> 30339
1.000	0.699	30339, 10331 -> 30032
1.000	0.699	30339, 30361 -> 30032
1.000	0.699	30339, 30327 -> 30032
1.000	0.699	30032, 30361 -> 30339
1.000	0.699	30339, 30316 -> 30032
1.000	0.699	30032, 30309 -> 30339
1.000	0.699	30339, 30303 -> 30032
1.000	0.699	30339, 30306 -> 30032
1.000	0.699	30032, 30314 -> 30339
1.000	0.699	30032, 30324 -> 30339
1.000	0.699	30339, 30319 -> 30032
1.000	0.699	30339, 30272 -> 30032
1.000	0.699	30339, 30363 -> 30032
1.000	0.699	30032, 30336 -> 30339
1.000	0.699	30032, 30315 -> 30339
1.000	0.699	30032, 30327 -> 30339
1.000	0.699	30032, 30344 -> 30339
1.000	0.699	30339, 30331 -> 30032
1.000	0.699	30339, 30309 -> 30032
1.000	0.699	30032, 30312 -> 30339
1.000	0.699	30032, 30313 -> 30339
1.000	0.699	30032, 30306 -> 30339
1.000	0.699	30339 -> 30032
1.000	0.699	30032, 30305 -> 30339
1.000	0.699	30032, 30307 -> 30339
1.000	0.699	30032, 30308 -> 30339
1.000	0.699	30032, 30311 -> 30339
1.000	0.699	30339, 30314 -> 30032
1.000	0.699	30339, 30315 -> 30032
1.000	0.699	30032, 30316 -> 30339
1.000	0.699	30032 -> 30339

2. <u>Type of crime location data rule mining:</u>
   - We got the dependency network. Here, we will select a historic location for illustration. We can see that if a crime happened in a healthcare and/or emergency location with/without another location type, then that crime could also happen in a historic location. Otherwise, the historic location had "2-way relationship" with other type of crime location.



Mining Model: v Assoc Seq Crime

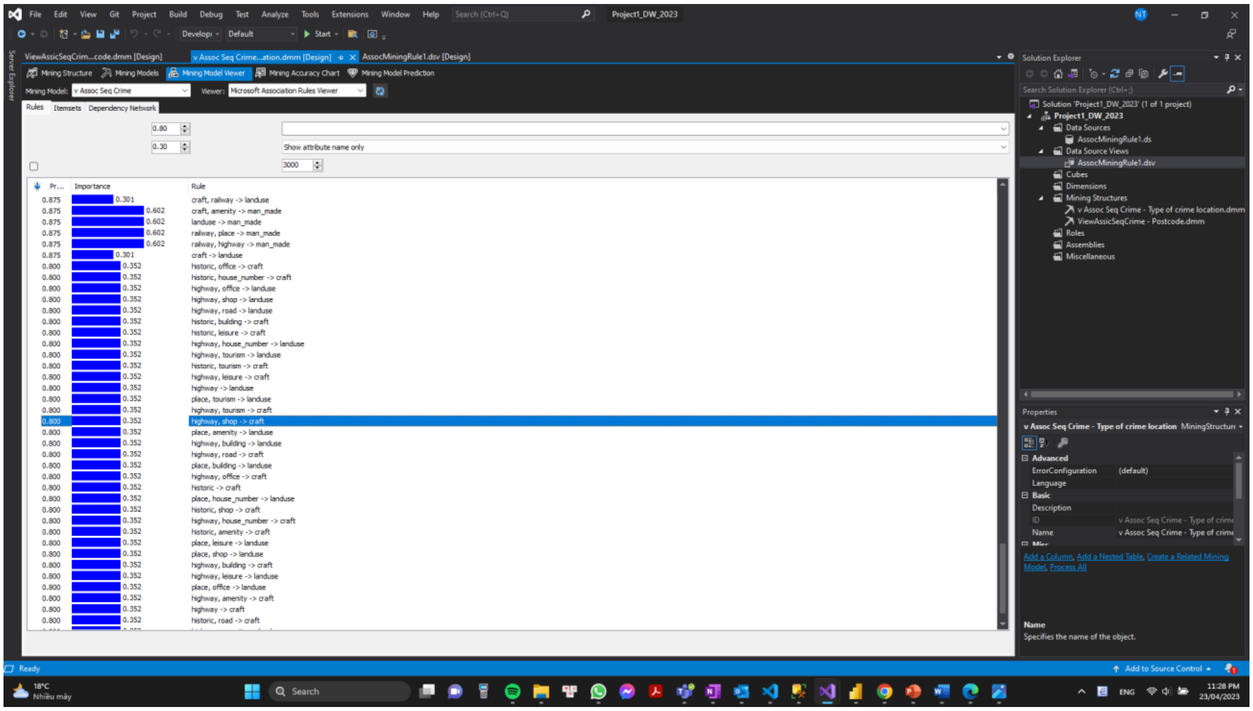
Viewer: Microsoft Association Rules Viewer

Show: Show attribute name and value

Select a node in the network to highlight its dependencies.

- Selected node
- This node predicts the selected node
- Predicts both ways
- Selected node predicts this node

- We got the rule mining model. We selected the probability of 80% and importance of 30%, then we found many rules. In the one highlighted in blue in the picture below, if a crime happened in a highway and in a shop, then there is 80% that crime would happen in a craft.



Pr...	Importance	Rule
0.875	0.301	craft, railway -> landuse
0.875	0.602	craft, amenity -> man_made
0.875	0.602	landuse -> man_made
0.875	0.602	railway, place -> man_made
0.875	0.602	railway, highway -> man_made
0.875	0.301	craft -> landuse
0.800	0.352	historic, office -> craft
0.800	0.352	historic, house_number -> craft
0.800	0.352	highway, office -> landuse
0.800	0.352	highway, shop -> landuse
0.800	0.352	highway, road -> landuse
0.800	0.352	historic, building -> craft
0.800	0.352	historic, leisure -> craft
0.800	0.352	highway, house_number -> landuse
0.800	0.352	highway, tourism -> landuse
0.800	0.352	historic, tourism -> craft
0.800	0.352	highway, leisure -> craft
0.800	0.352	highway -> landuse
0.800	0.352	place, tourism -> landuse
0.800	0.352	highway, tourism -> craft
0.800	0.352	highway, shop -> craft
0.800	0.352	place, amenity -> landuse
0.800	0.352	highway, building -> landuse
0.800	0.352	highway, road -> craft
0.800	0.352	place, building -> landuse
0.800	0.352	highway, office -> craft
0.800	0.352	historic -> craft
0.800	0.352	place, house_number -> landuse
0.800	0.352	historic, shop -> craft
0.800	0.352	highway, house_number -> craft
0.800	0.352	historic, amenity -> craft
0.800	0.352	place, leisure -> landuse
0.800	0.352	place, shop -> landuse
0.800	0.352	highway, building -> craft
0.800	0.352	highway, leisure -> landuse
0.800	0.352	place, office -> landuse
0.800	0.352	highway, amenity -> craft
0.800	0.352	highway -> craft
0.800	0.352	historic, road -> craft

C. Top k rule explanation:

K is the association rule which is most frequent or common. For example, if k=3 and we have minimum of confidence of 90%, then we can get top 3 rules with a confidence equal or higher than 90%