MSIS 5663 Data Warehousing

OLAP Cube Design & Data Mining

Submitted by:

Akshay Arora(A20101619)

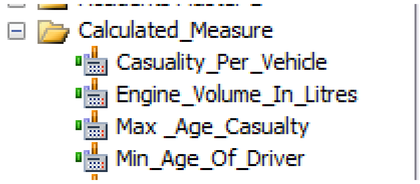
Ankur Bansal (A20079925)

Reno Rajan Christy (A20082578)

Viswesh Muralitharan (A20095678)

Contents

1. NAMED CALCULATION / MEASURES:

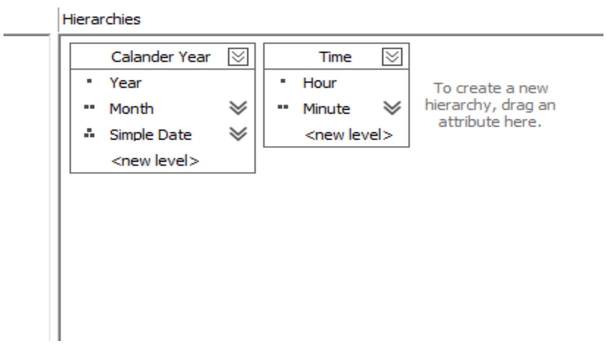


We created above 4 named Measures as seen above in the Team5_MDM_New.

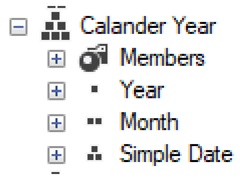
- Casualty per vehicle - Number of casualties / number of vehicles involved in the accident
- Engine Volume in Liters - Calculate the Engine capacity in Liters
- Minimum Age of Driver – calculates minimum age of driver
- Maximum Age of Casualty– calculates maximum age of causality

2. DATE HIERARCHIES:

- Hierarchies are relationships among the attributes of a dimension mostly a one to many relationship.

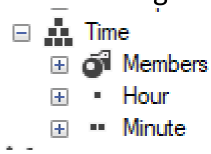


Calendar Year Hierarchy:
The Calendar year hierarchy has three levels namely Year (Level 1), Month(Level 2) and Simple date(Level 3) where the month level member is being identified as a combination of Year and month keys. Similarly simple date is identified in relationship through a combination of month and Simple date key. The relationships are one to many.



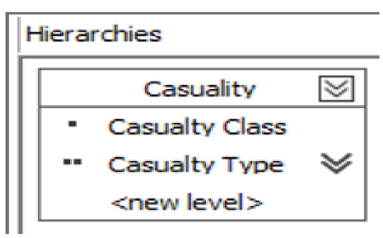
Time Hierarchy:

Time Hierarchy has two levels namely hour (Level 1) and minute (Level 2) where the minute level is being identified as a combination of hour and minute.



3. CASUALTY HIERARCHY:

Casualty Hierarchy has two levels namely Casualty Class (Level1) and Casualty Type (Level 2) where the Casualty Type is identified as a combination of Casualty Class and Casualty Type with one to many relationship.



4. PARTITIONS AND AGGREGATIONS:

PARTITIONS:

- Partitions are used by Microsoft SQL Server Analysis to manage and store data and aggregations for a measure group in a cube.

- One of the advantages is that partitions can be processed separately and can use different partitions.
- Only current partition needs to be processed when the current information is added to the cube; Processing smaller amount of data will improve processing performance by decreasing processing time.

Our Partitions:

Since we have eleven years we decided to create 3 partitions as below:

1) Accident Master 2005-2008
2) Accident Master 2009-2012
3) Accident Master 2013-2015

	Partition Name	Source	Estimated Rows	Storage Mode	Aggregation Design
1	Accidents Master 2005-2008	SELECT [dbo].[Accidents_Master].[Accident_key],[dbo].[Accidents_Master].[Accident...	4650859	MOLAP	AggregationDesign30PercentPerfor...
2	Accidents Master 2009-2012	SELECT [dbo].[Accidents_Master].[Accident_key],[dbo].[Accidents_Master].[Accident...	0	MOLAP	AggregationDesign30PercentPerfor...
3	Accidents Master 2013-2015	SELECT [dbo].[Accidents_Master].[Accident_key],[dbo].[Accidents_Master].[Accident...	0	MOLAP	AggregationDesign30PercentPerfor...

Justification:

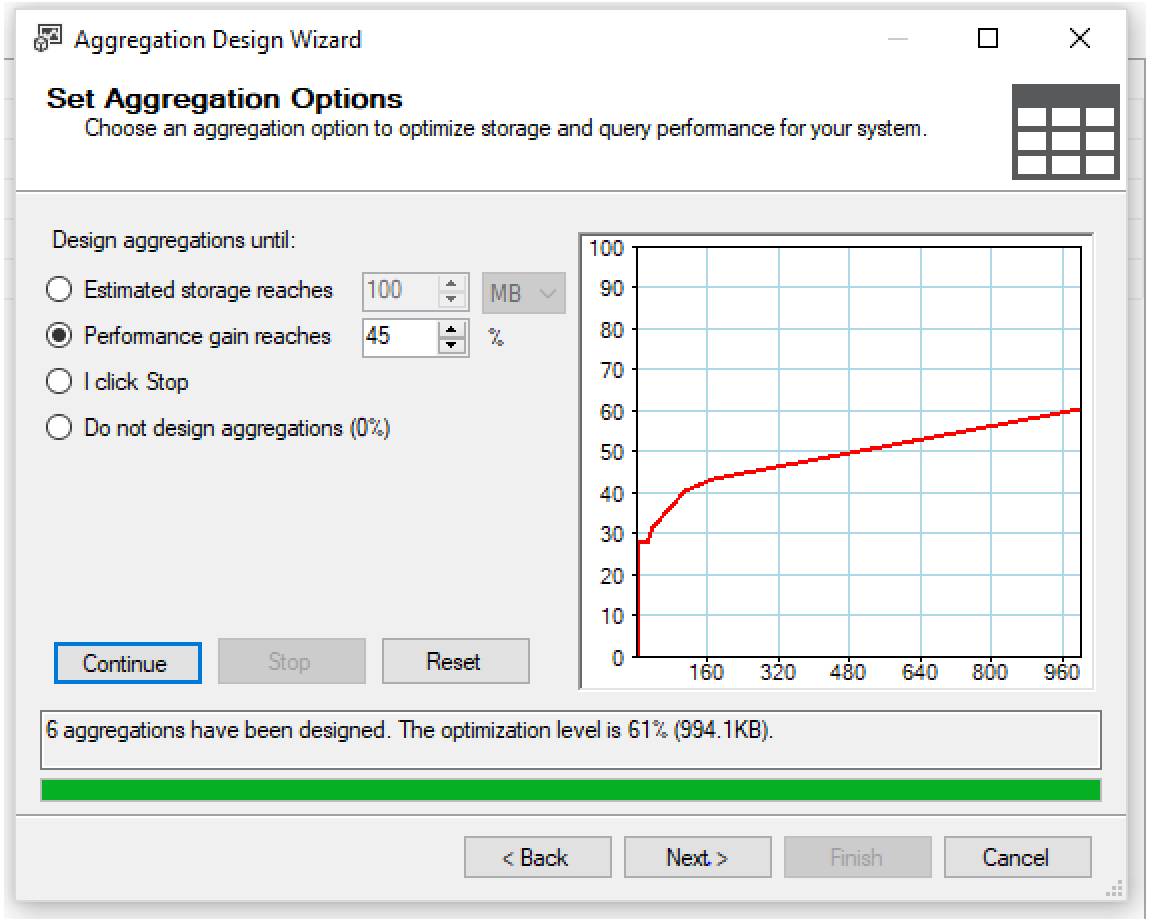
As we can see, our Accidents Database contains 4.6 million records of historical data in yearly basis. So, the Partitions are decided based on the fact that data is being assumed to be refreshed after 3-4 years and the last partition being for the recent years increases query performance as most the business queries is being expected to be done on recent data. Hence the above partitions would improve the performance from the subjective stand point.

AGGREGATIONS:

- An aggregation is a pre-calculated summary of facts from leaf cells representing summarization of measure group at certain granularity of dimensions. Aggregations occur during processing of the cube and aggregations have to be stored separately.
- 100% aggregations is not even necessary as some aggregations can be calculated from other pre-calculated aggregations.

Our Aggregation design and Justification:

Aggregation Design wizard provides options for us to specify storage and percentage constraints on the algorithm to achieve a satisfactory tradeoff between query response time and storage requirements. I.e. as aggregations percentage increases storage requirements decreases, but, the query response time decreases and vice versa. Therefore, we have used 45% performance gain to our cube partitions as an optimal performance solutions as shown below.



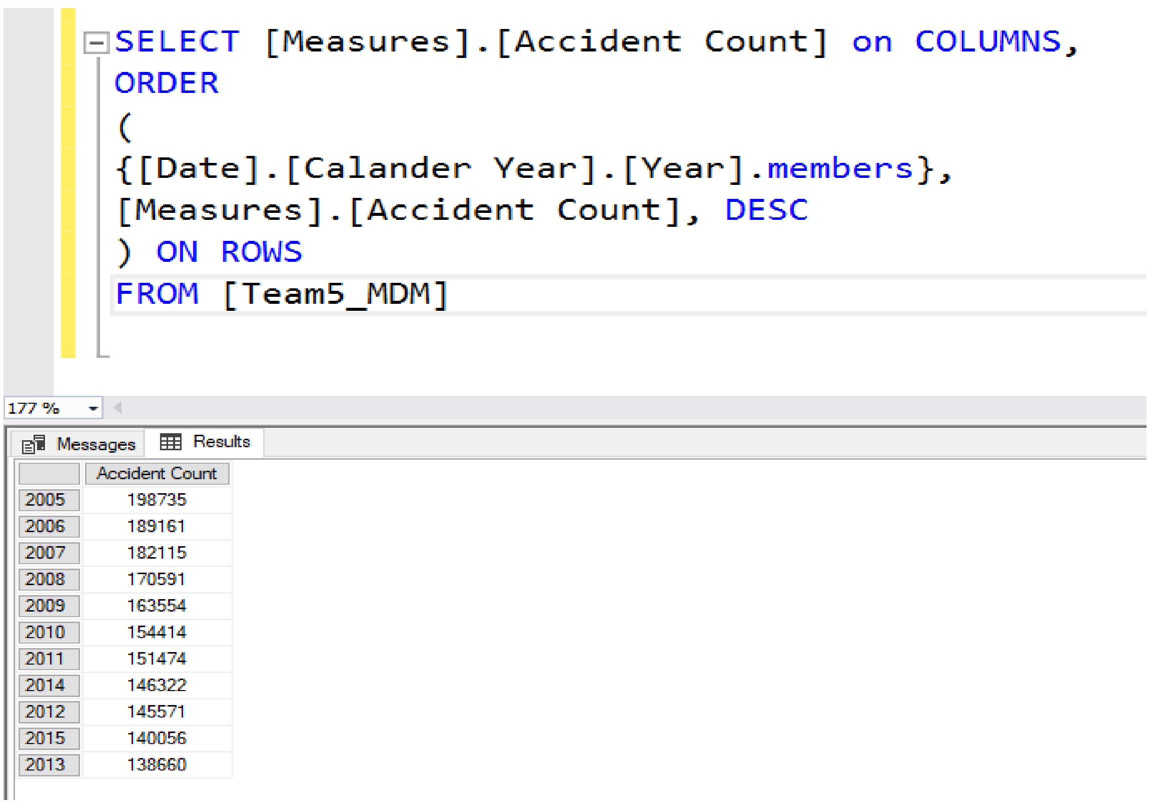
Since, aggregations go hand in hand with optimal performance we decided to put a performance gain of 45% which aggregates four aggregations and keeps it in the cube. Also, the storage requirement for above aggregation is less and hence it wouldn't be an overhead for cube memory storage.

5. REPORTING WITH MDX QUERIES:

As we are dealing with the accidents data, we are mostly analyzing the accident related aspects through the following reporting queries. There can be many reporting queries that can be developed beyond the below ones, but, we have considered the queries that would give a quick insight to the user about the accidents data.

1. Display Accident Count for each year in descending order of Accident count

```
SELECT [Measures].[Accident Count] on COLUMNS,
ORDER
(
{[Date].[Calander Year].[Year].members},
[Measures].[Accident Count], DESC
) ON ROWS
FROM [Team5_MDM]
```



```
SELECT [Measures].[Accident Count] on COLUMNS,
ORDER
(
{[Date].[Calander Year].[Year].members},
[Measures].[Accident Count], DESC
) ON ROWS
FROM [Team5_MDM]
```

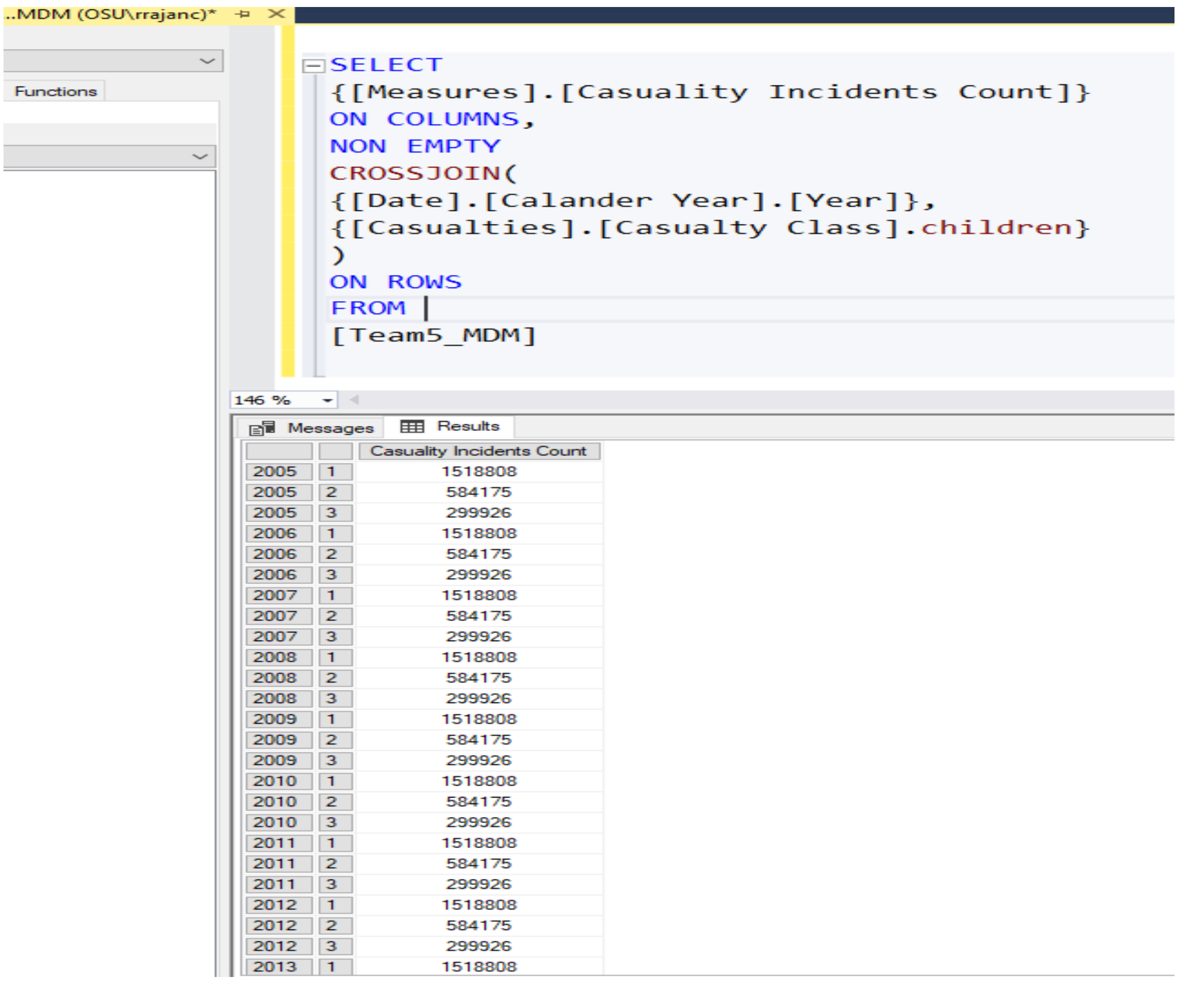
177 %

Messages Results

	Accident Count
2005	198735
2006	189161
2007	182115
2008	170591
2009	163554
2010	154414
2011	151474
2014	146322
2012	145571
2015	140056
2013	138660

2. Display the total casualty count for each year

```
SELECT
{[Measures].[Casuality Incidents Count]}
ON COLUMNS,
NON EMPTY
CROSSJOIN(
{[Date].[Calander Year].[Year]},
{[Casualties].[Casualty Class].children}
)
ON ROWS
FROM
[Team5_MDM]
```



		Casuality Incidents Count
2005	1	1518808
2005	2	584175
2005	3	299926
2006	1	1518808
2006	2	584175
2006	3	299926
2007	1	1518808
2007	2	584175
2007	3	299926
2008	1	1518808
2008	2	584175
2008	3	299926
2009	1	1518808
2009	2	584175
2009	3	299926
2010	1	1518808
2010	2	584175
2010	3	299926
2011	1	1518808
2011	2	584175
2011	3	299926
2012	1	1518808
2012	2	584175
2012	3	299926
2013	1	1518808

3. Display the accident count for each year where the severity of the injury is maximum.

```
SELECT
   [Measures].[Accident Count] ON COLUMNS,
   {([Date].[Calander Year].[Year])}*
   {EXCEPT(
```

```
   {[Casualties].[Casualty Severity].children},

   {[Casualties].[Casualty Severity].&[2],[Casualties].[Casualty Severity].&[3]})} ON
ROWS
FROM
    [Team5_MDM]
```

```
SELECT
   [Measures].[Accident Count] ON COLUMNS,
   {([Date].[Calander Year].[Year])}*
   {EXCEPT(
   {[Casualties].[Casualty Severity].children},

   {[Casualties].[Casualty Severity].&[2],[Casualties].[Casualty Severity].&[3]})} ON ROWS
FROM
    [Team5_MDM]
```

46 %

Messages | Results

		Accident Count
2005	1	2913
2006	1	2926
2007	1	2714
2008	1	2341
2009	1	2057
2010	1	1731
2011	1	1797
2012	1	1637
2013	1	1608
2014	1	1658
2015	1	1616

4. Display the accident count for each year of male sex driver [Assumed Males as 1]

```
SELECT [Measures].[Accident Count] ON 0,
[Date].[Calander Year].[Year].members*EXISTS(
[Vehicles].[Sex Of Driver].[Sex Of Driver].MEMBERS,
 {[Vehicles].[Sex Of Driver].&[1]}
) ON 1
FROM [Team5_MDM]
```

```
SELECT [Measures].[Accident Count] ON 0,
[Date].[Calander Year].[Year].members*EXISTS(
[Vehicles].[Sex Of Driver].[Sex Of Driver].MEMBERS,
 {[Vehicles].[Sex Of Driver].&[1]}
) ON 1
FROM [Team5_MDM]
```

46 %

Messages Results

		Accident Count
2005	1	166583
2006	1	157856
2007	1	151814
2008	1	141065
2009	1	134914
2010	1	126770
2011	1	124117
2012	1	119364
2013	1	113739
2014	1	120522
2015	1	115539

5. Show for each year, show the Calendar Month with the highest accident count where accident caused by skidding and turning excluding the invalid values

```
WITH SET [Months With High Accident Counts Per Year] AS
Generate( [Date].[Calander Year].[Year].MEMBERS,
TopCount(
Descendants( [Date].[Calander Year].CurrentMember, [Date].[Calander Year].[Month],SELF ),
1,
[Measures].[Accident Count] ))

SELECT
NON EMPTY
{[Months With High Accident Counts Per Year] * [Measures].[Accident Count]}
ON 0,
NON EMPTY
except(
{[Vehicles].[Skidding And Overturning].AllMEMBERS}, [Vehicles].[Skidding And
Overturning].&[-1]
)
ON 1
FROM
[Team5_MDM]
```

```
WITH SET [Months With High Accident Counts Per Year] AS
Generate( [Date].[Calander Year].[Year].MEMBERS,
TopCount(
Descendants( [Date].[Calander Year].CurrentMember, [Date].[Calander Year].[Month],SELF ), 1,
[Measures].[Accident Count] ))

SELECT
NON EMPTY
{[Months With High Accident Counts Per Year] * [Measures].[Accident Count]}
ON 0,
NON EMPTY
except(
{[Vehicles].[Skidding And Overturning].AllMEMBERS}, [Vehicles].[Skidding And Overturning].&[-1]
)
ON 1
FROM
[Team5_MDM]
```

	11	11	11	10	11	11	10	11	10	10	7
	Accident Count	Accident Count	Accident Count	Accident Count	Accident Count	Accident Count	Accident Count	Accident Count	Accident Count	Accident Count	Accident Count
All	18747	17397	16559	15684	15473	14544	13748	13305	13322	13450	12771
0	16704	15584	14818	14164	13840	12961	12579	12076	12215	12381	11722
1	3764	3210	2892	2553	2834	2572	1802	1843	1754	1715	1541
2	687	618	605	531	542	522	453	412	382	420	370
3	10	10	9	14	14	14	7	5	8	9	12
4	8	4	3	8	5	7	5	2	2	2	1
5	464	455	415	382	353	312	339	312	343	352	416

6. Display the accident count for the year 2005 and for the month of February in the same year.

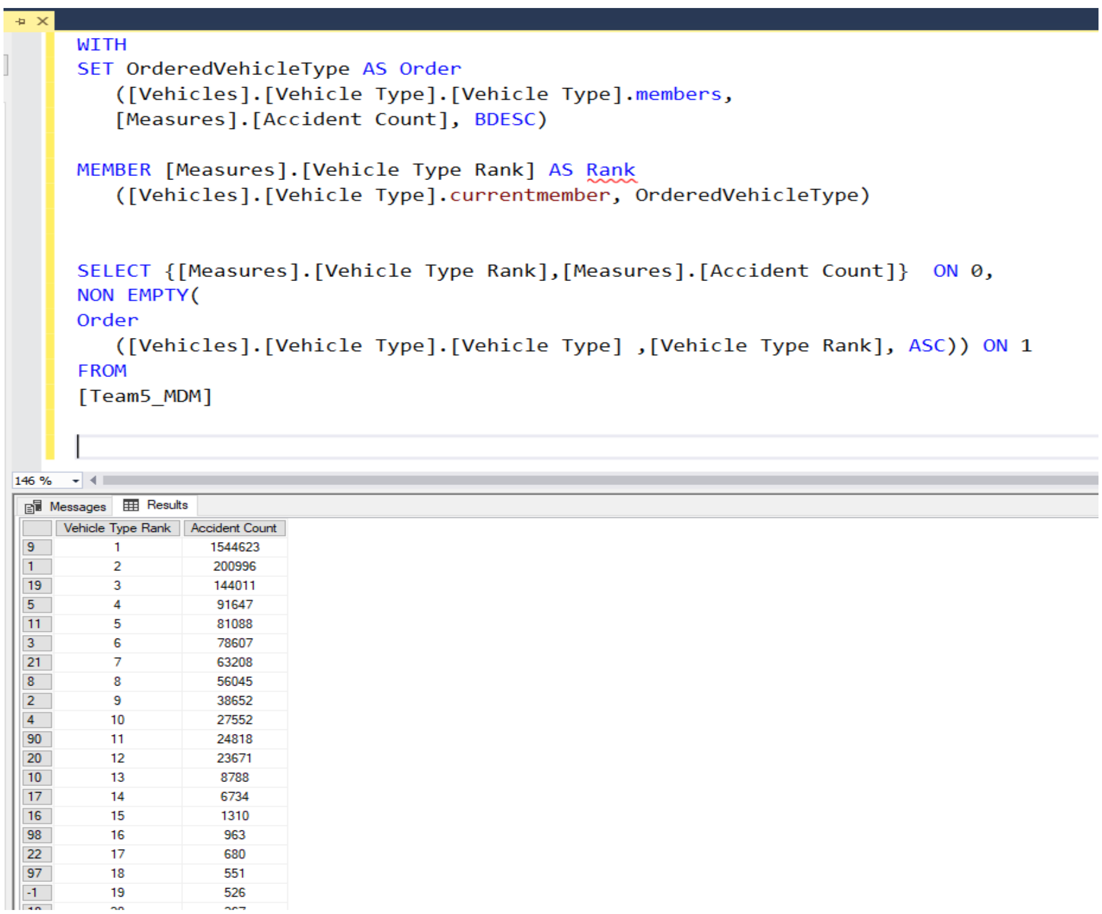
```
SELECT
{[Measures].[Accident Count]} ON COLUMNS,
{
(Ancestors( [Date].[Calander Year].[Month].&[2]&[2005],0)),
(Ancestors([Date].[Calander Year].[Month].&[2]&[2005],1))
}
ON ROWS
FROM
[Team5_MDM]
```

```
SELECT
{[Measures].[Accident Count]} ON COLUMNS,
{
(Ancestors( [Date].[Calander Year].[Month].&[2]&[2005],0)),
(Ancestors([Date].[Calander Year].[Month].&[2]&[2005],1))
}
ON ROWS
FROM
[Team5_MDM]
```

	Accident Count
2	14521
2005	198735

7. Display the Vehicle types, Vehicle Type Rank and Accident Count for all Vehicle Types in order from highest Accident Count to lowest.

```
WITH
SET OrderedVehicleType AS Order
   ([Vehicles].[Vehicle Type].[Vehicle Type].members,
   [Measures].[Accident Count], BDESC)

MEMBER [Measures].[Vehicle Type Rank] AS Rank
   ([Vehicles].[Vehicle Type].currentmember, OrderedVehicleType)


SELECT {[Measures].[Vehicle Type Rank],[Measures].[Accident Count]}  ON 0,
NON EMPTY(
Order
   ([Vehicles].[Vehicle Type].[Vehicle Type] ,[Vehicle Type Rank], ASC)) ON
1
FROM
[Team5_MDM]
```



```
WITH
SET OrderedVehicleType AS Order
   ([Vehicles].[Vehicle Type].[Vehicle Type].members,
   [Measures].[Accident Count], BDESC)

MEMBER [Measures].[Vehicle Type Rank] AS Rank
   ([Vehicles].[Vehicle Type].currentmember, OrderedVehicleType)


SELECT {[Measures].[Vehicle Type Rank],[Measures].[Accident Count]}  ON 0,
NON EMPTY(
Order
   ([Vehicles].[Vehicle Type].[Vehicle Type] ,[Vehicle Type Rank], ASC)) ON 1
FROM
[Team5_MDM]
```

146 %

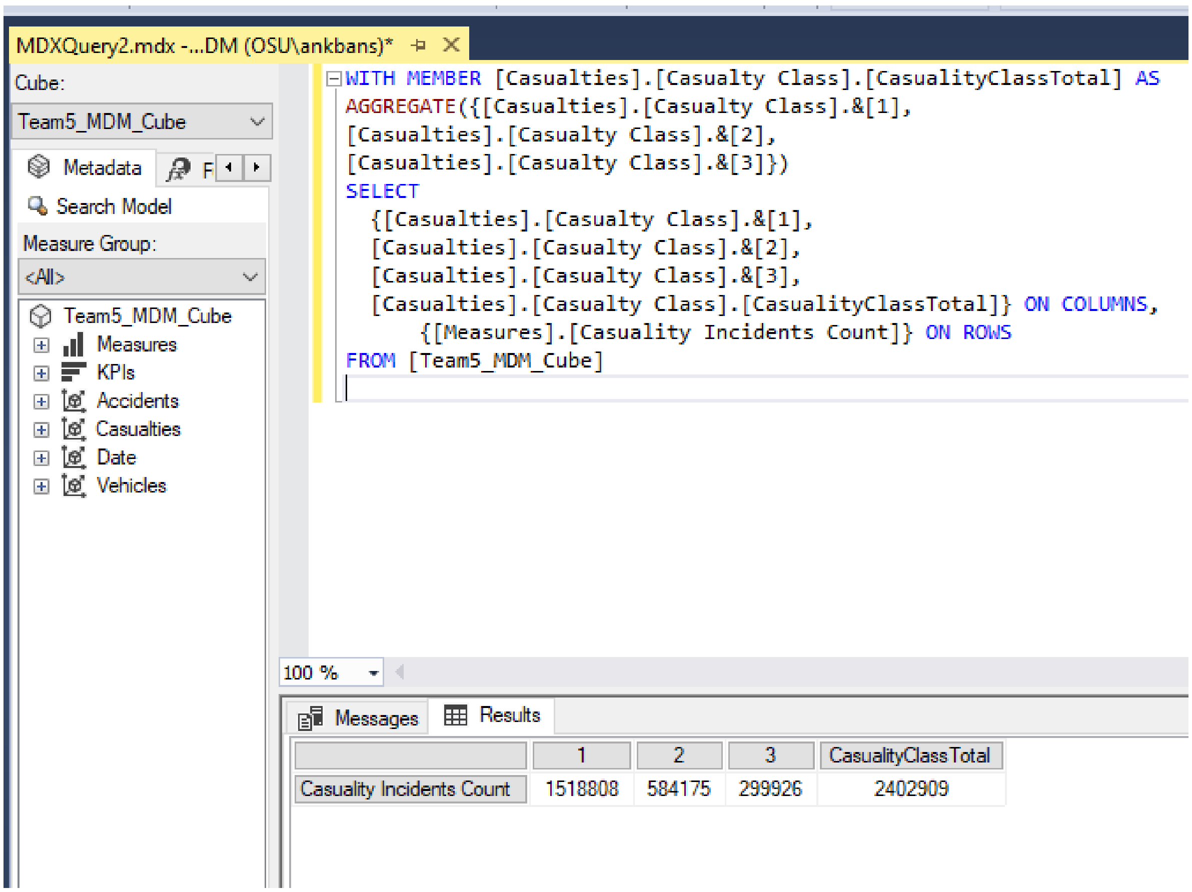
Messages | Results

	Vehicle Type Rank	Accident Count
9	1	1544623
1	2	200996
19	3	144011
5	4	91647
11	5	81088
3	6	78607
21	7	63208
8	8	56045
2	9	38652
4	10	27552
90	11	24818
20	12	23671
10	13	8788
17	14	6734
16	15	1310
98	16	963
22	17	680
97	18	551
-1	19	526

8. Display the Causality Counts for each class and then their aggregated total

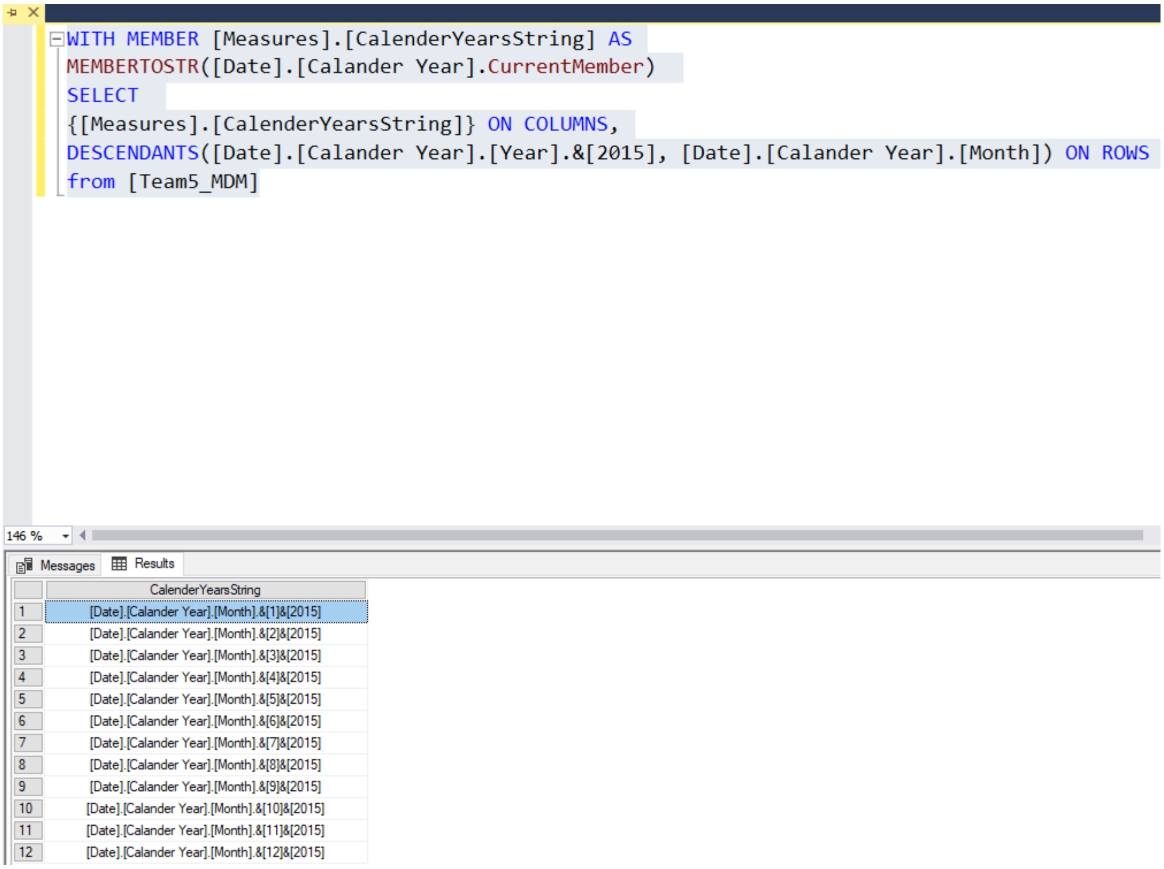
```
WITH MEMBER [Casualties].[Casualty Class].[CasualityClassTotal] AS
```

```
AGGREGATE({[Casualties].[Casualty Class].&[1],
[Casualties].[Casualty Class].&[2],
[Casualties].[Casualty Class].&[3]})
SELECT
{[Casualties].[Casualty Class].&[1],
[Casualties].[Casualty Class].&[2],
[Casualties].[Casualty Class].&[3],
[Casualties].[Casualty Class].[CasualityClassTotal]} ON COLUMNS,
{[Measures].[Casuality Incidents Count]} ON ROWS
FROM [Team5_MDM]
```



9. Display the Multidimensional Expressions formatted string of Year 2015 and its months using the created Date hierarchy

```
WITH MEMBER [Measures].[CalenderYearsString] AS
MEMBERTOSTR([Date].[Calander Year].CurrentMember)
SELECT
{[Measures].[CalenderYearsString]} ON COLUMNS,
DESCENDANTS([Date].[Calander Year].[Year].&[2015], [Date].[Calander Year].[Month]) ON
ROWS FROM [Team5_MDM]
```



10. For each vehicle type, for each date with speed limits of 50 and 70, display their respective non empty accident counts

```
SELECT [Measures].[Accident Count] ON COLUMNS,
EXTRACT(
NONEMPTY
(
{[Vehicles].[Vehicle Type].members
*
[Date].[Date].[Date].MEMBERS}
*
{[Accidents].[Speed Limit].&[50],
[Accidents].[Speed Limit].&[70]}
*
{[Measures].[Accident Count]}
)
,[Vehicles].[Vehicle Type],[Date].[Date],[Accidents].[Speed Limit]
)
ON ROWS
FROM [Team5_MDM]
```

```
SELECT [Measures].[Accident Count] ON COLUMNS,
EXTRACT(
NONEMPTY
(
{[Vehicles].[Vehicle Type].members
*
[Date].[Date].[Date].MEMBERS}
*
{[Accidents].[Speed Limit].&[50],
[Accidents].[Speed Limit].&[70]}
*
{[Measures].[Accident Count]}
)
,[Vehicles].[Vehicle Type],[Date].[Date],[Accidents].[Speed Limit]
)
ON ROWS
FROM [Team5_MDM]
```

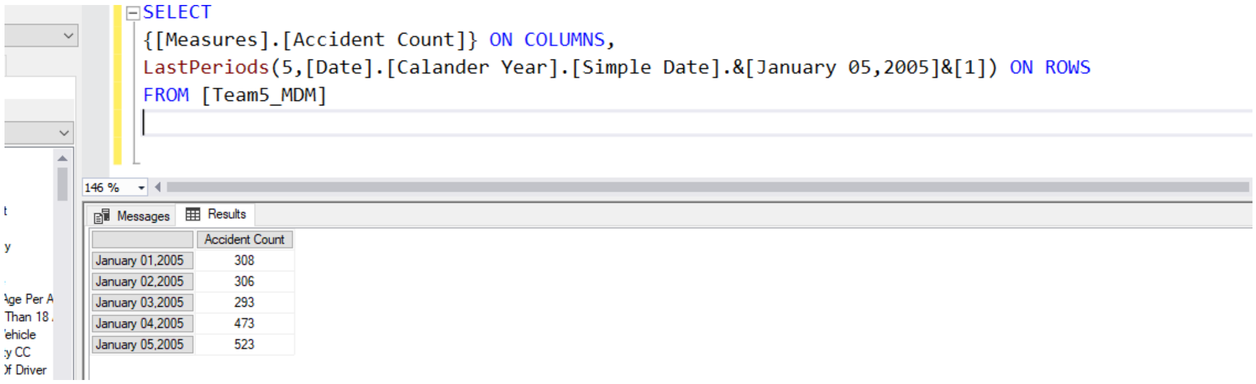
146 %

Messages | Results

			Accident Count
All	2005-01-01	50	6
All	2005-01-01	70	20
All	2005-01-02	50	13
All	2005-01-02	70	35
All	2005-01-03	50	5
All	2005-01-03	70	23
All	2005-01-04	50	14
All	2005-01-04	70	34
All	2005-01-05	50	13
All	2005-01-05	70	46
All	2005-01-06	50	14
All	2005-01-06	70	49
All	2005-01-07	50	15
All	2005-01-07	70	38
All	2005-01-08	50	12
All	2005-01-08	70	42
All	2005-01-09	50	10
All	2005-01-09	70	31
All	2005-01-10	50	14
All	2005-01-10	70	40

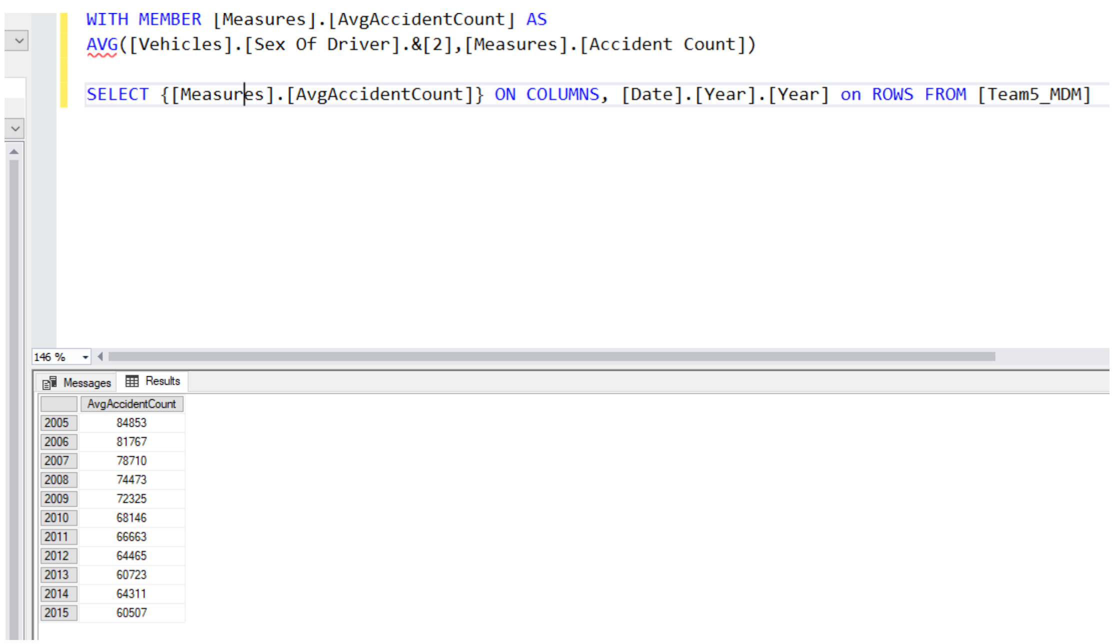
11. Display the Accident Count for first five days of January 2005.

```
SELECT
{[Measures].[Accident Count]} ON COLUMNS,
LastPeriods(5,[Date].[Calander Year].[Simple Date].&[January 05,2005]&[1]) ON ROWS
FROM [Team5_MDM]
```



12. Display the Average Accident Count of Female Drivers for each year

```
WITH MEMBER [Measures].[AvgAccidentCount] AS
AVG([Vehicles].[Sex Of Driver].&[2],[Measures].[Accident Count])

SELECT {[Measures].[AvgAccidentCount]} ON COLUMNS, [Date].[Year].[Year] on ROWS FROM
[Team5_MDM]
```



DELIVERABLE 2 : DATA MINING & SUMMARY OF FINDINGS

Objectives:

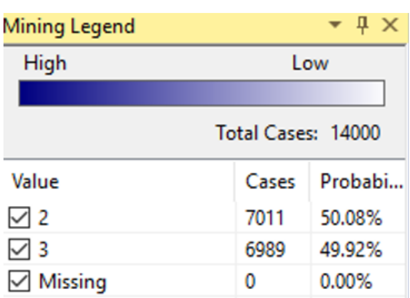
Our intent is to predict the accident severity, specifically interested in the Major injury severity of the person in an accident

- To uncover major features and factors that contribute to severe injury in accidents in USA
- To understand most risk prone factors that lead to severe injury
- To identify relevant factors that favors minor injury over severe injury

6. DATA MINING MODELS

Our modelling approach was to randomly sample the data to have a balanced dataset which contains almost an equal number of majority and minority classes, here in this case being Severe/Major injury(Accident Severity =2) and Minor injury (Accident Severity =3). The data balancing was obtained using an equal number of records from both classes to train the model as shown in the queries below using Union function.

Below we can see the balanced trained data which is 70% of the total data (20000 records)



CREATING MINING STRUCTURE & MODELS FOR DATA MINING

```
CREATE MINING STRUCTURE [Team5_MDM DMX]
(
  [Accident Index] LONG KEY,
  [Accident Severity] LONG DISCRETE,
  [Carriageway Hazards] TEXT DISCRETE,
  [First Road Class] TEXT DISCRETE,
  [Junction Control] TEXT DISCRETE,
  [Junction Detail] TEXT DISCRETE,
  [Light Conditions] TEXT DISCRETE,
  [Ped Cross Human] TEXT DISCRETE,
  [Ped Cross Physical] TEXT DISCRETE,
  [Police Officer Attend] TEXT DISCRETE,
  [Road Surface Conditions] TEXT DISCRETE,
  [Road Type] TEXT DISCRETE,
  [Second Road Class] TEXT DISCRETE,
  [Special Conditions At Site] TEXT DISCRETE,
  [Speed Limit] TEXT DISCRETE,
 [Urban Rural] TEXT DISCRETE,
 [Weather Conditions] TEXT DISCRETE
)
WITH HOLDOUT (30 PERCENT)
```

Adding Mining Model with Decision Tree Model:-

```
ALTER MINING STRUCTURE [Team5_MDM DMX]
ADD MINING MODEL [Decision Tree DMX]
(
 [Accident Index],
 [Accident Severity] PREDICT,
 [Carriageway Hazards],
  [First Road Class],
  [Junction Control],
  [Junction Detail],
  [Light Conditions],
  [Ped Cross Human],
  [Ped Cross Physical],
  [Police Officer Attend],
  [Road Surface Conditions],
```

```
  [Road Type],
  [Second Road Class],
  [Special Conditions At Site],
  [Speed Limit],
 [Urban Rural],
 [Weather Conditions]
) USING Microsoft_Decision_Trees
WITH DRILLTHROUGH
```

Adding Mining Model with Association Model:-

```
ALTER MINING STRUCTURE [Team5_MDM DMX]
ADD MINING MODEL [Association DMX]
(
 [Accident Index],
 [Accident Severity] PREDICT,
 [Carriageway Hazards],
  [First Road Class],
  [Junction Control],
  [Junction Detail],
  [Light Conditions],
  [Ped Cross Human],
  [Ped Cross Physical],
  [Police Officer Attend],
  [Road Surface Conditions],
  [Road Type],
  [Second Road Class],
  [Special Conditions At Site],
  [Speed Limit],
 [Urban Rural],
 [Weather Conditions]
) USING Microsoft_Association_Rules
WITH DRILLTHROUGH

DELETE FROM MINING STRUCTURE [Team5_MDM DMX]



ALTER MINING STRUCTURE [Team5_MDM DMX]
ADD MINING MODEL [Clustering DMX]
```

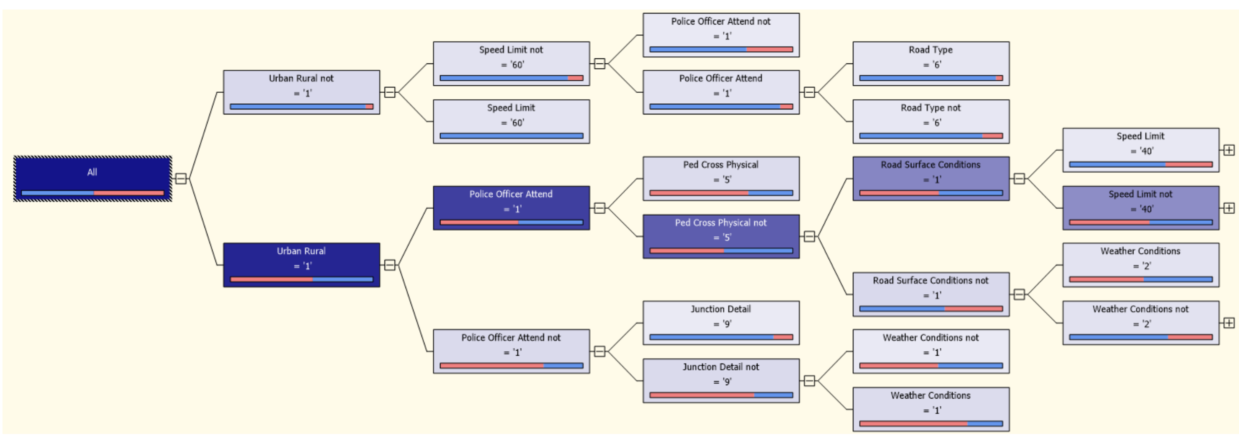
```
(
[Accident Index],
  [Accident Severity] PREDICT,
  [Carriageway Hazards],
   [First Road Class],
   [Junction Control],
   [Junction Detail],
   [Light Conditions],
   [Ped Cross Human],
   [Ped Cross Physical],
   [Police Officer Attend],
   [Road Surface Conditions],
   [Road Type],
   [Second Road Class],
   [Special Conditions At Site],
   [Speed Limit],
  [Urban Rural],
  [Weather Conditions]
) USING Microsoft_Clustering
WITH DRILLTHROUGH

ALTER MINING STRUCTURE [Team5_MDM DMX]
ADD MINING MODEL [Neural_Network DMX]
(
[Accident Index],
  [Accident Severity] PREDICT,
  [Carriageway Hazards],
   [First Road Class],
   [Junction Control],
   [Junction Detail],
   [Light Conditions],
   [Ped Cross Human],
   [Ped Cross Physical],
   [Police Officer Attend],
   [Road Surface Conditions],
   [Road Type],
   [Second Road Class],
   [Special Conditions At Site],
   [Speed Limit],
  [Urban Rural],
  [Weather Conditions]
) USING Microsoft_Neural_Network
```

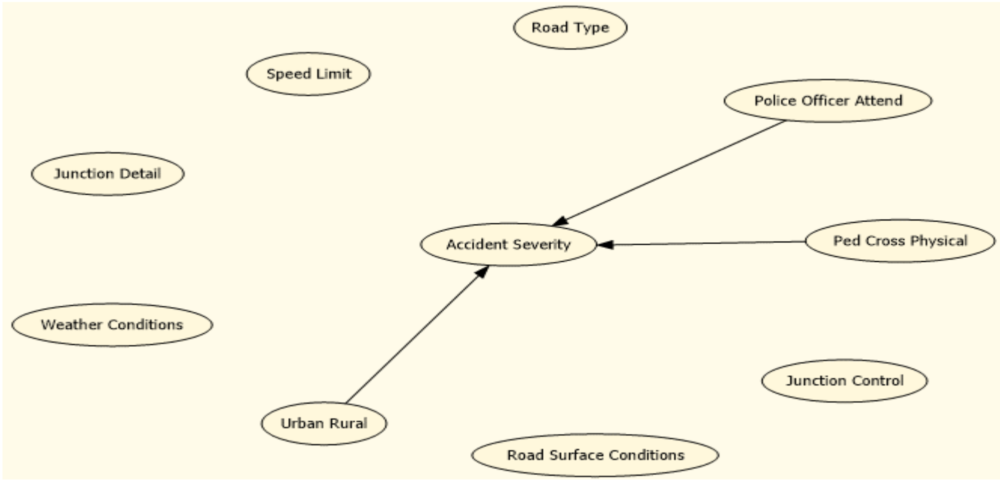
Insert Records in the Mining Structure:

```
INSERT INTO MINING STRUCTURE [Team5_MDM DMX]
(
   [Accident Index],
   [Accident Severity],
   [Carriageway Hazards],
   [First Road Class],
   [Junction Control],
   [Junction Detail],
   [Light Conditions],
   [Ped Cross Human],
   [Ped Cross Physical],
   [Police Officer Attend],
   [Road Surface Conditions],
   [Road Type],
   [Second Road Class],
   [Special Conditions At Site],
   [Speed Limit],
  [Urban Rural],
  [Weather Conditions]
)
OPENQUERY([UK Accidents Database],
   'SELECT TOP 10000 [Accident_Index],
   [Accident_Severity],
   [Carriageway_Hazards],
   [First_Road_Class],
   [Junction_Control],
   [Junction_Detail],
   [Light_Conditions],
   [Ped_Cross_Human],
   [Ped_Cross_Physical],
   [Police_Officer_Attend],
   [Road_Surface_Conditions],
   [Road_Type],
   [Second_Road_Class],
   [Special_Conditions_At_Site],
   [Speed_Limit],
  [Urban_Rural],
  [Weather_Conditions]
    FROM DimAccidents where [Accident_Severity]=2
UNION
SELECT TOP 10000 [Accident_Index],
   [Accident_Severity],
```

```
[Carriageway_Hazards],
[First_Road_Class],
[Junction_Control],
[Junction_Detail],
[Light_Conditions],
[Ped_Cross_Human],
[Ped_Cross_Physical],
[Police_Officer_Attend],
[Road_Surface_Conditions],
[Road_Type],
[Second_Road_Class],
[Special_Conditions_At_Site],
[Speed_Limit],
[Urban_Rural],
[Weather_Conditions]
FROM DimAccidents where [Accident_Severity]=3')
```

DECISION TREE MODEL RESULTS & EVALUATION

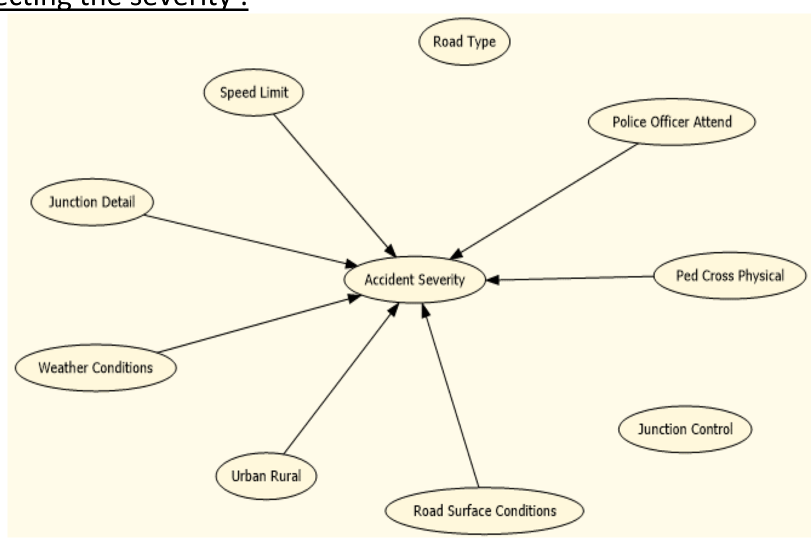




Dependency network reveals that in general the Urban/Rural location owned followed by Police officer attending the scene and Ped Cross Physical are the primary determinants for identifying whether the accident would be severe or not.

The decision tree indicate that when the locality of the incident is Rural (I.e Urban Rural not is 1) then the larger segment of the accident in comparison to the total is more likely to be severe. The second factor that profoundly affects the severity is the Speed limit of 60 miles. However when the locality setting is Urban in nature, the police officer attending the case is being considered as the next impactful variable in best classifying the severity of the model

<u>Factors least affecting the severity :</u>

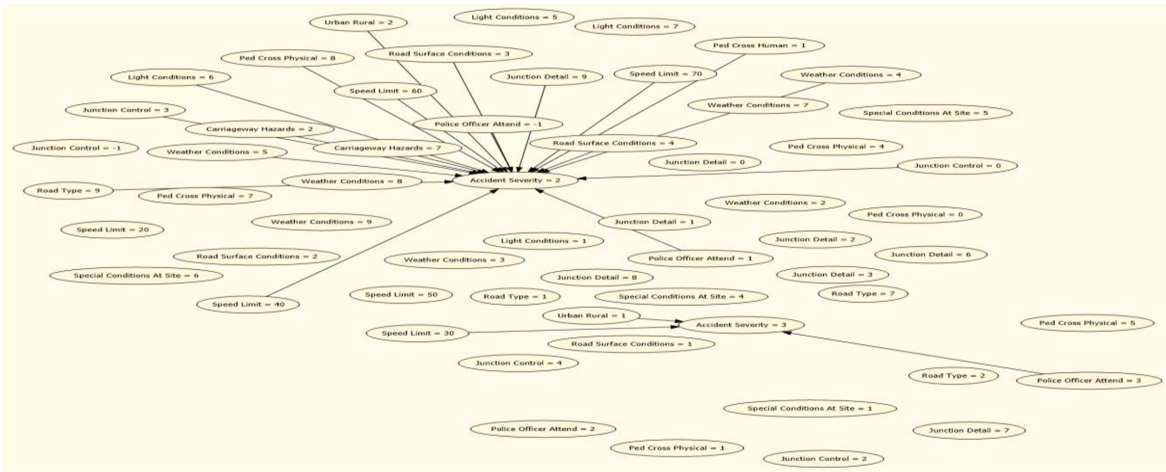


Surprisingly the factors that least affect the severity turns out to be the Road type and junction control as shown above.

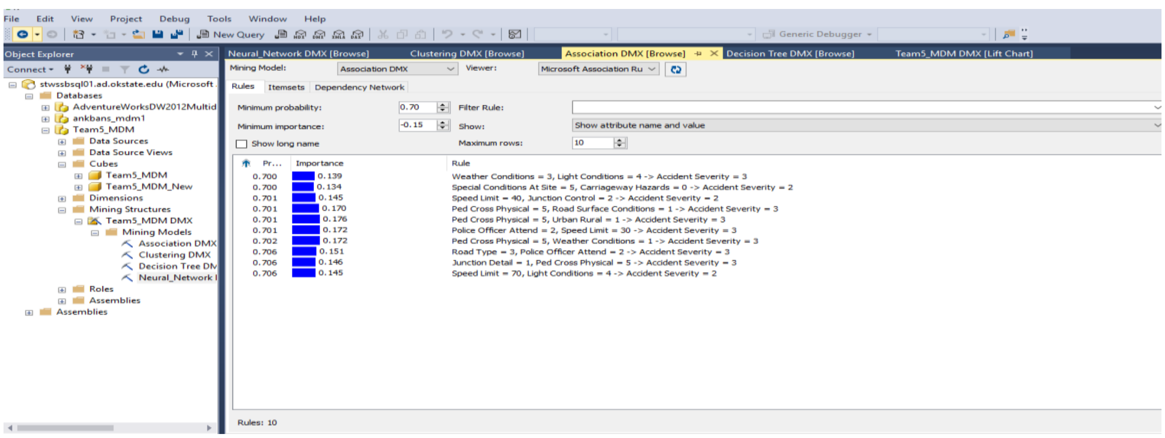
The main factors that influence decision trees are: -

Urban Rural, Police Officer Attend, Ped Cross Physical, Speed Limit

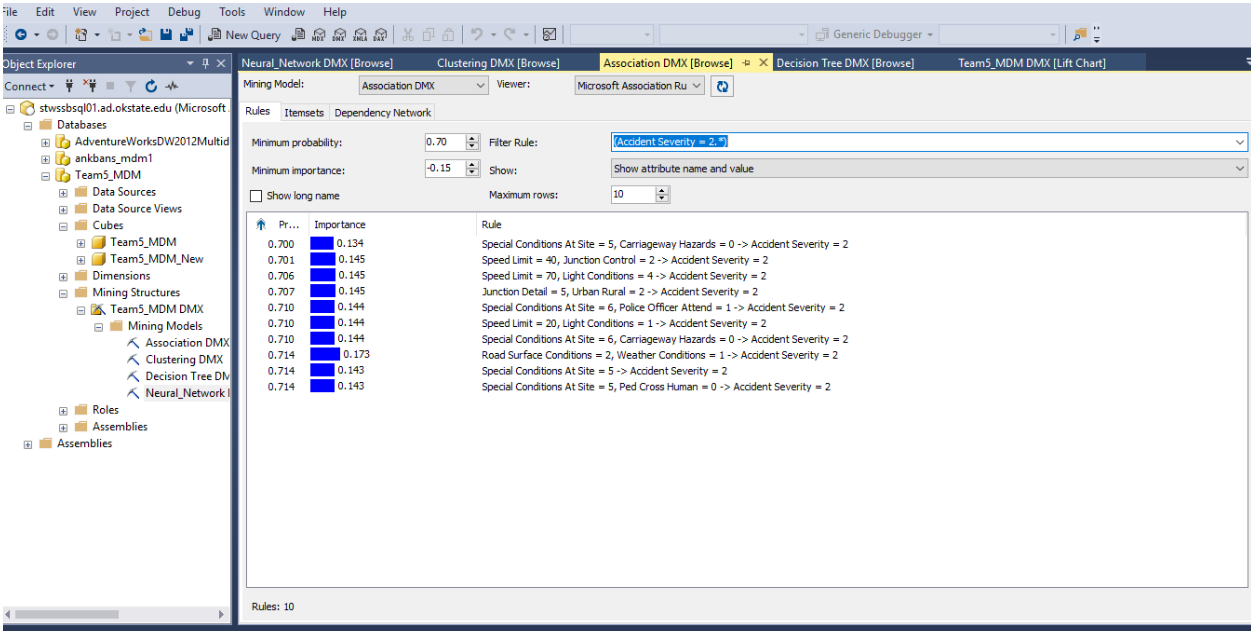
ASSOCIATION MODEL EVALUATION



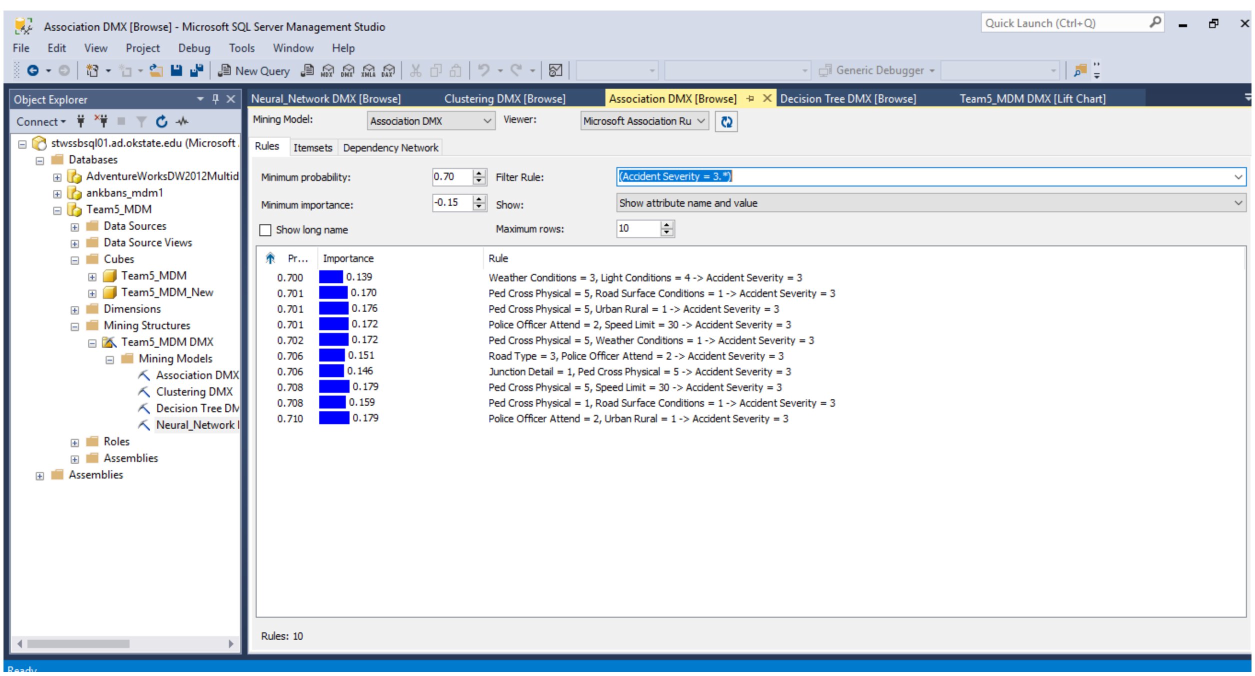
Factors that affect the Association model in general based on dependency network is:
Urban Rural=2,Speed Limit=60,Light Condition=6,Junction Control =0, Weather Condition=5



Based on importance we can conclude that Ped cross physical being 5 and Urban Rural being 1 mostly results in severity minor.

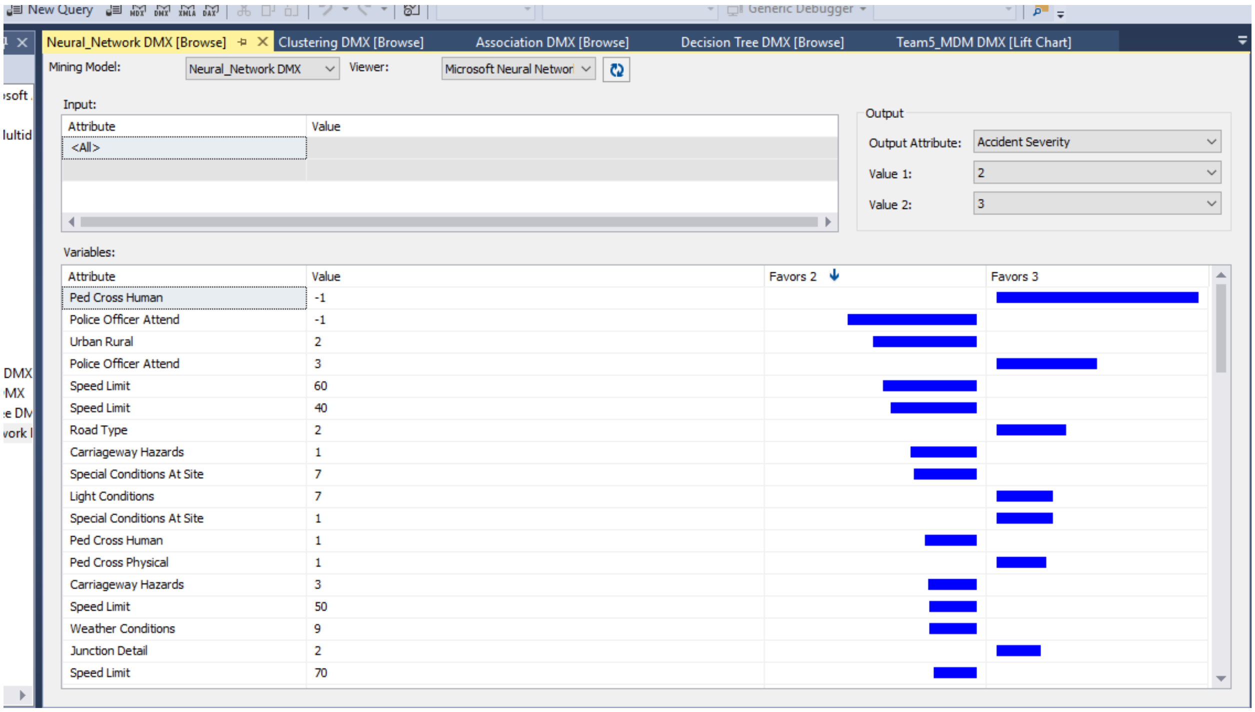


For the Severe injury accidents (I.e. 2), The road surface conditions and weather conditions has high impact on the severity. When road is wet and damp along with a fine weather would result in high severity accidents.



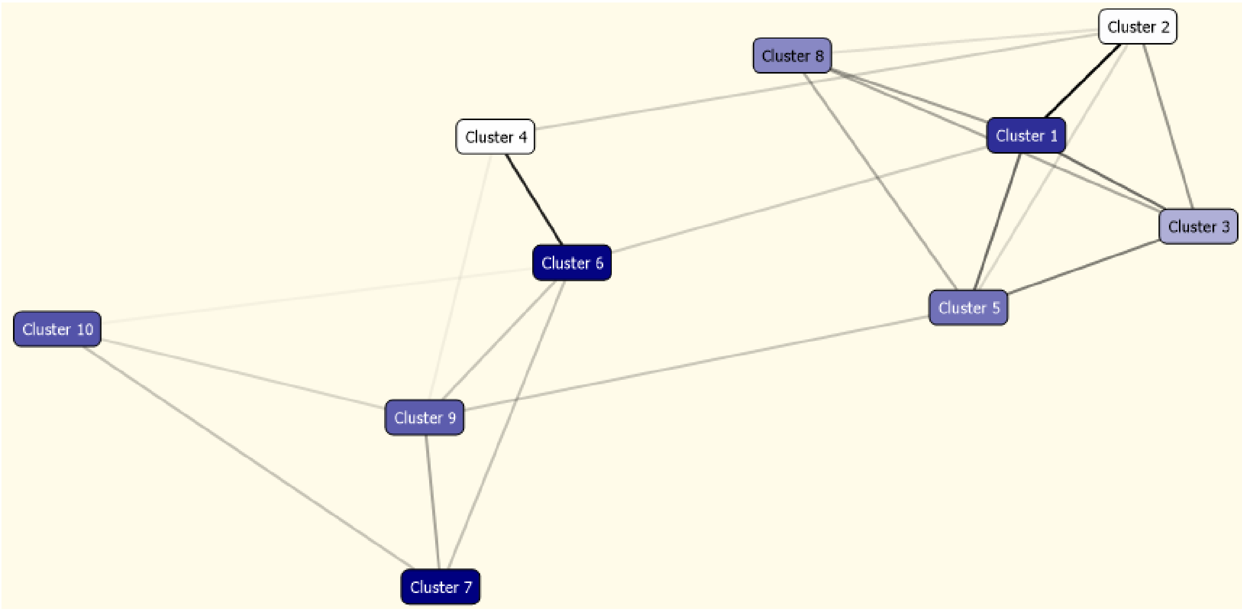
The high importance of 0.179 reveals that wherever there is a pedestrian phase in the traffic signal with a lower speed limit of 30 results in low severity accident incidents.

NEURAL NETWORK MODEL



The factors that mostly identify severe injury would be Police officer attending the event being unknown followed by Rural locality, speed limit favours Severe accident injury over minor injury. On the other hand Pedestrian cross places with police presence areas without accidents and One way roads favours minor accident cases over severe ones.

CLUSTERING MODEL



Major factors influence:-

Urban Rural, Speed Limit, Light Conditions, Junction Control

For Accident Severity = 2, Cluster 7 is the most dominant as compared to other clusters which has 98% of Accident Severity=2, followed by cluster 6, cluster 1.

Here, it is interesting to note that the light conditions=6 which means Darkness-No lightening or almost zero visibility has the maximum percentage in cluster 7 which means this light condition category highly influenced towards major severity in the accident. Also, the analysis shows that speed Limit=60 was also a contributing factors towards major severity in an accident. Also, junction control =0 which means Not near the junction is also contributing towards major severity of an accident. Overall, the results of cluster analysis coincides with other model results and provides us similar influencing factors.

7 . PERFORMANCE COMPARISON BETWEEN MODELS

CONFUSION MATRIX:

Counts for Decision Tree DMX on Accident Severity			
	Predicted	2 (Actual)	3 (Actual)
	2	1322	308
	3	1667	2703
Counts for Association DMX on Accident Severity			
	Predicted	2 (Actual)	3 (Actual)
	2	2005	999
	3	984	2012
Counts for Clustering DMX on Accident Severity			
	Predicted	2 (Actual)	3 (Actual)
	2	1634	728
	3	1355	2283
Counts for Neural_Network DMX on Accident Severity			
	Predicted	2 (Actual)	3 (Actual)
	2	2723	2244
	3	266	767

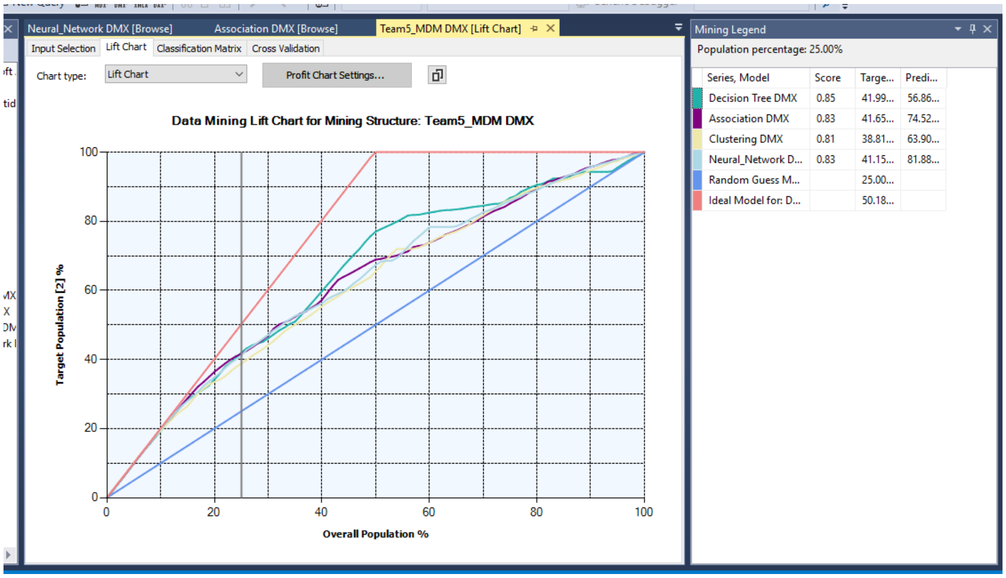
Model	**Accuracy %**
Decision Tree	**67.08**
Association Rule	66.95
Clustering	65.28
Neural Network	58.17

The Decision Tree seems to have the highest accuracy in general compared to all other models.

COMPARISON BASED ON LIFT SCORE

Series, Model	Score	Targe...	Predi...
Decision Tree DMX	0.85	41.99...	56.86...
Association DMX	0.83	41.65...	74.52...
Clustering DMX	0.81	38.81...	63.90...
Neural_Network D...	0.83	41.15...	81.88...
Random Guess M...		25.00...	
Ideal Model for: D...		50.18...	

The Decision tree model has an overall best lift score and accuracy in comparison to other models and hence can be selected as the best model for predicting the accident severity class in general.



- The chart shows that if you target (for example) 25% of the population (20000), I.e. 5000 people; with the
    - the random guess model will correctly identify 25% of all severe injuries(0.25* 10000= 2500) in the population
    - Our decision tree prediction model will correctly identify 41.99% of all severe injuries (0.4199* 10000= 4199) in the population
    - ideal line (perfect prediction model), you will correctly identify 50.18% of all severe injuries (0.5018* 10000=5018) in the population

8. BEST PERFORMING ALGORITHM

COMPARISON BASED ON ACCURACY, SENSITIVITY & SPECIFICITY:

Model	**Accuracy %**	**Sensitivity %**	**Specificity %**
Decision Tree	**67.08**	44.22	**89.77**
Association Rule	66.95	67.08	66.82
Clustering	65.28	54.67	75.82
Neural Network	58.17	**91.10**	25.47

Since the Neural network model has the best sensitivity we would ideally select the same as the best performing model as our intent is primarily focused on the Severe /Fatal Injury (2) cases. The sensitivity of the Neural network model is 91%. However Decision Tree model can be selected as a good model to predict the dichotomous variable in general as well to predict the minor injury category as its specificity is high. The overall accuracy of the Decision tree is 67% and specificity is 89%.

9.SUMMARY OF FINDINGS & RECOMMENDATIONS

Major features and factors that contribute to severe injury in accidents in USA

- Significant factors that lead to severe injury is as below:
  - Visibility factors like Darkness with limited or no lighting.
  - Factors affecting vehicle control like rain, slippery roads in combination with high speed of the vehicle.
  - Adverse weather conditions like affecting vehicle maneuver, visibility and control like snow and rainy conditions which result in slippery roads.
  - Rural areas are more prone to severe accidents probably owing to the lack of infrastructure like street lights, road conditions coupled with lack of junction control.
- Factors that favor minor injury over severe injury
  - Police surveillance/reporting seems to have a positive affect on reducing the severity of the accident.
  - Low speed of vehicle results in minor injury over severe injury.
  - Presence of pedestrians results in reduced injury severity possibly attributed to the fact that drivers are cautious in pedestrian commuted areas.
  - In General factors that least affect the severity is the road type and the junction detail as discovered through dependency diagrams in many models.

RECOMMENDATIONS

- Proper infrastructure in the rural areas like improved lighting etc. could possibly reduce injury severity.
- Restrictive speed limits on adverse weather and light conditions is recommended.
- Cautioned driving in wet and slippery roads especially in rain and icy roads.
- In Rural areas speed limits should be restricted to less than 60 for reducing accident severity.