DATA WAREHOUSING
MSIS 5663
TERM PROJECT – SPRING 2019

Team 7
Roshith Elangovan- **A20161234**
Krithika Ganesh Kumar- **A20162108**
Anudeep Subraveti- **A20168818**

CONTENT

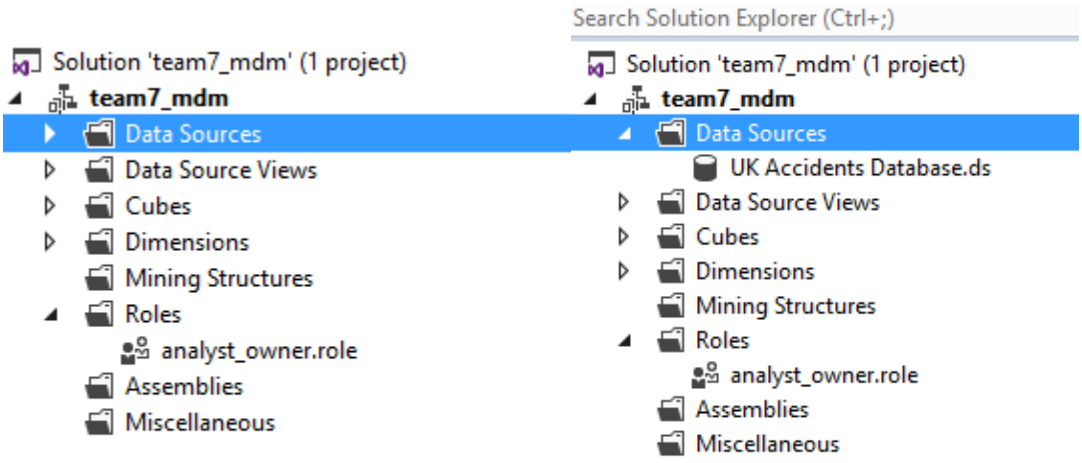
Introduction

The dataset given is “UK_Accidents_database”, is used in relational database format. It consists of 4,650,859 records and four dimensions namely Date, Vehicles, Causalities and Accidents.

The following steps have been followed:

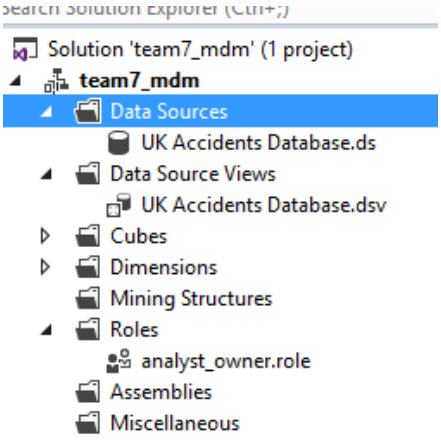
Step 1: Establishing Connection:

- A data Connection is made to UK_Accidents_database on the server (stwssbsql01.ad.okstate.edu)
- We selected NewProject>Import from Server from Analysis Server under Business Intelligence
- We added the data source

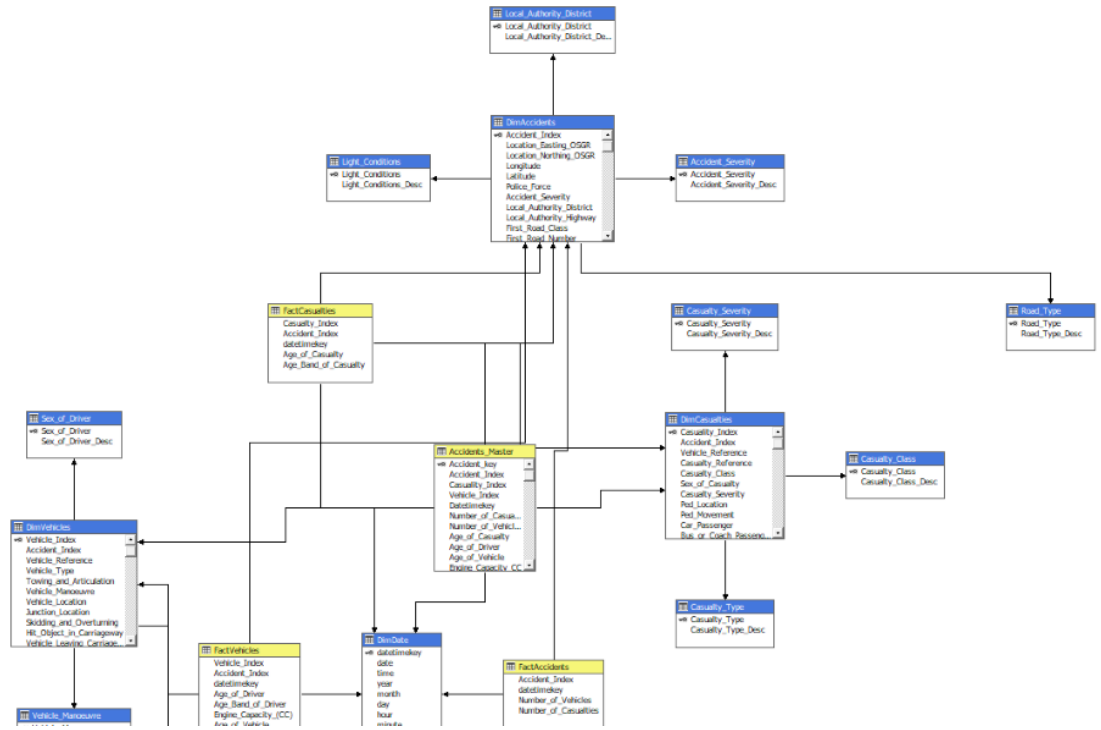


Step 2: Creating Data Source View:

- We created a Data Source View of the UK_Accidents_database which will serve as a disconnected data source for the project



We checked the data source view diagram



Step 3: Creating Named Calculations:

We created the following named calculations:

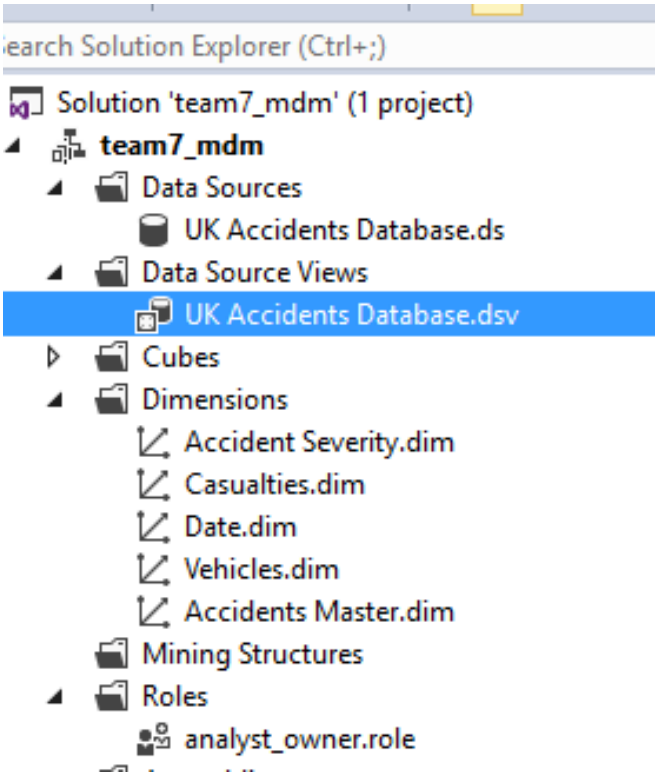
Named Calculation	Description
Police_Officer_Attended_Desc	Description of Police officer who attended the scene

Junc_Control_Desc	Description of the Junction at which the accident took place
Journey_Purpose_Desc	Journey of the passenger in the vehicle during the time of accident
DateofAccident	Date at which the accident took place in the format Month Date, Year

Step 4:Creating Dimensions:

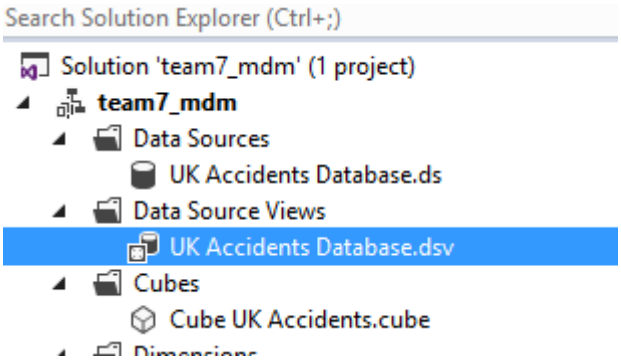
Next step is to create Dimensions .

We created the following Dimensions,



Step 5:Creating and deploying the cube:

Since the dimensions are intact , next step in the process is to create cube.



Step 6:Creating calculated measures:

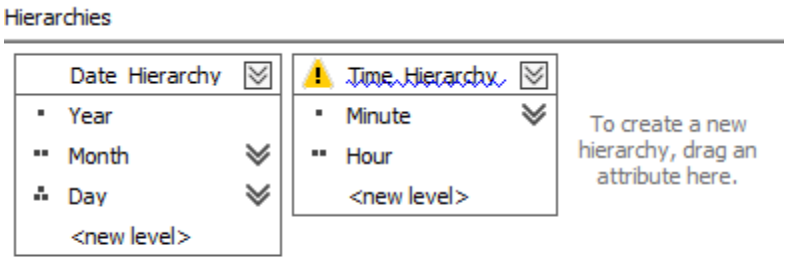
We created the following calculated measures:

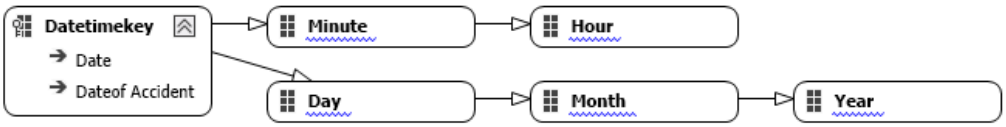
Measures	Description	Formula
[Average Age of Casualties]	This measures gives the average age of the casualties	[Measures].[Age Of Casualty]/[Measures].[Number Of Casualties]
[Driver to Casualties Ratio]	This gives the ratio between driver's age and casualty	[Measures].[Age Of Vehicle]/[Measures].[Number Of Vehicles]
Minimum Number of Casualties	This gives the minimum number of casulaties	
Engine_Capacity_CC Distinct Count	This gives the distinct count of Engine capacity that will help in determining the fatality of the accident	

Step 7:Creating Hierarchies:

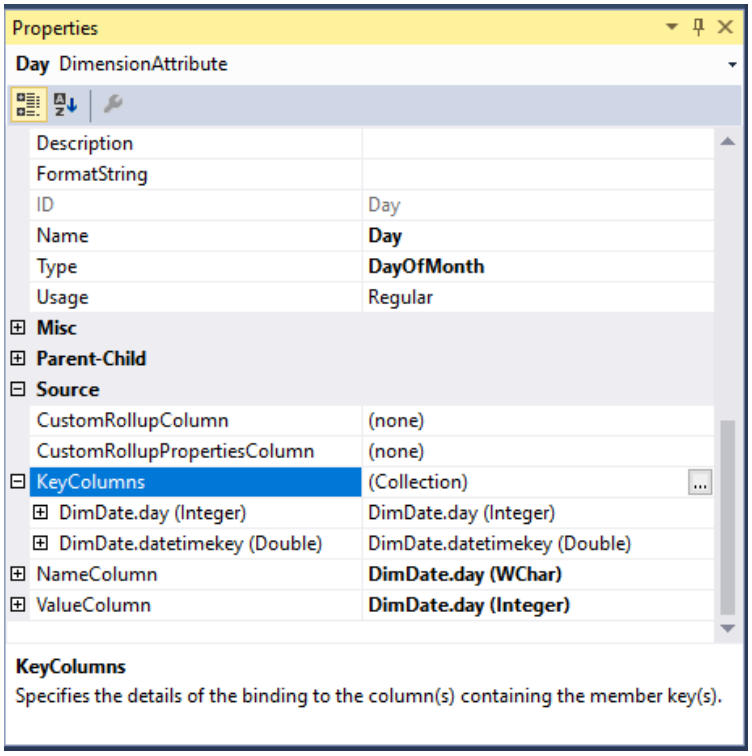
We created hierarchies in two dimensions:

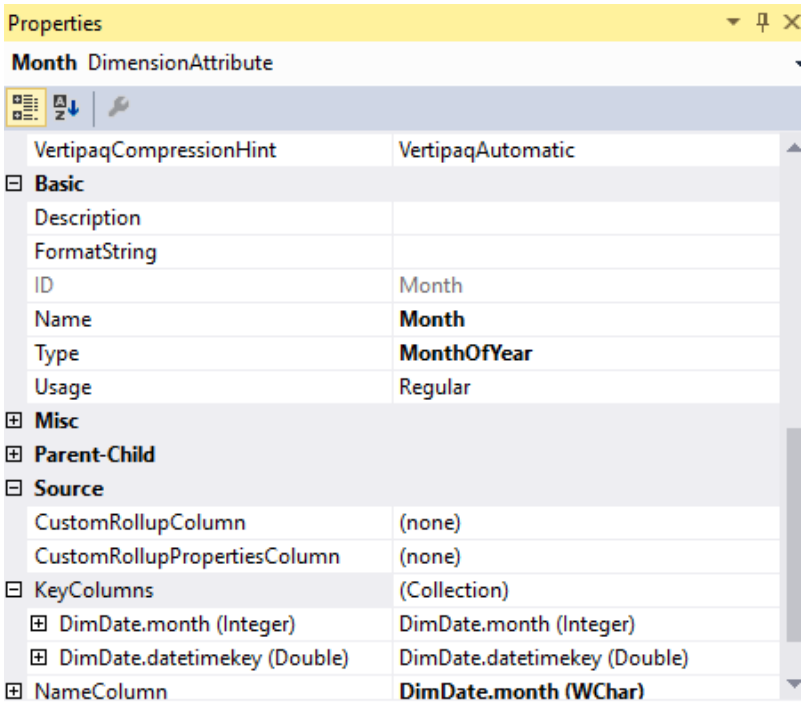
(i) Date Dimension:

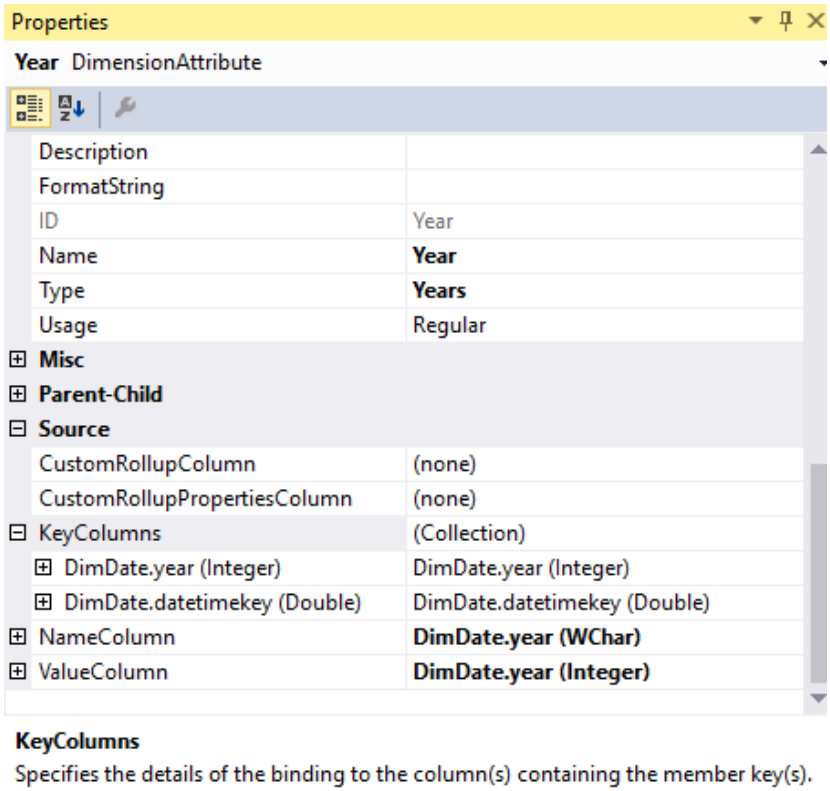




The properties of the attributes are as follows,

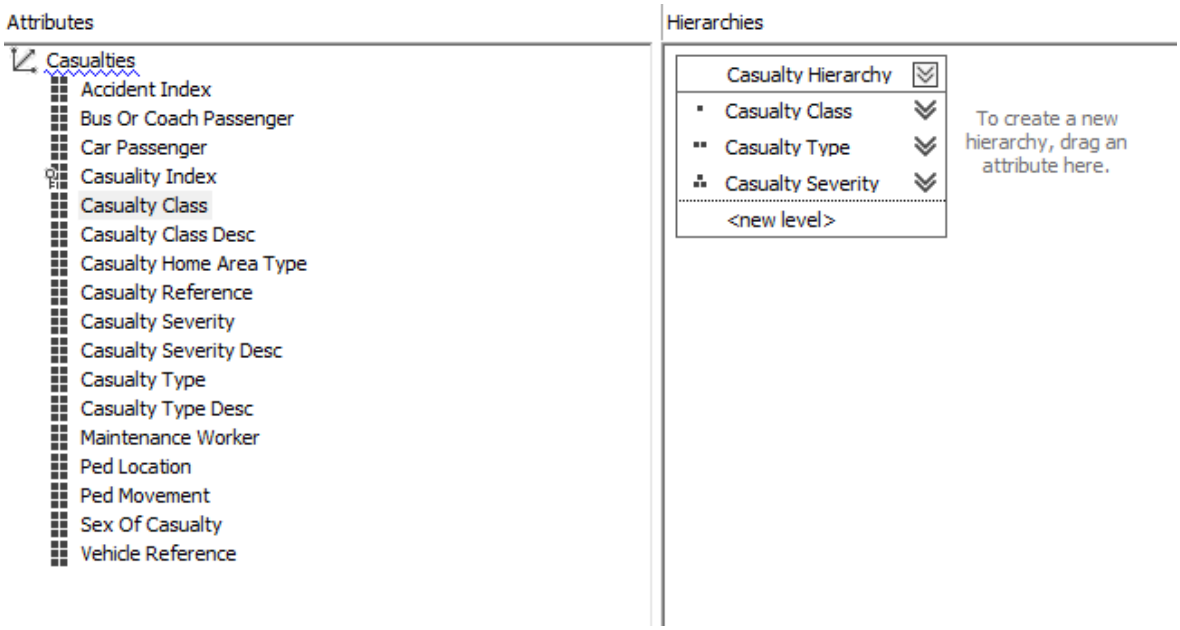



Properties
Month DimensionAttribute
VertipaqCompressionHint
VertipaqAutomatic
Basic
Description
FormatString
ID
Month
Name
Month
Type
MonthOfYear
Usage
Regular
Misc
Parent-Child
Source
CustomRollupColumn
(none)
CustomRollupPropertiesColumn
(none)
KeyColumns
(Collection)
DimDate.month (Integer)
DimDate.month (Integer)
DimDate.datetimekey (Double)
DimDate.datetimekey (Double)
NameColumn
DimDate.month (WChar)
KeyColumns
Specifies the details of the binding to the column(s) containing the member key(s).

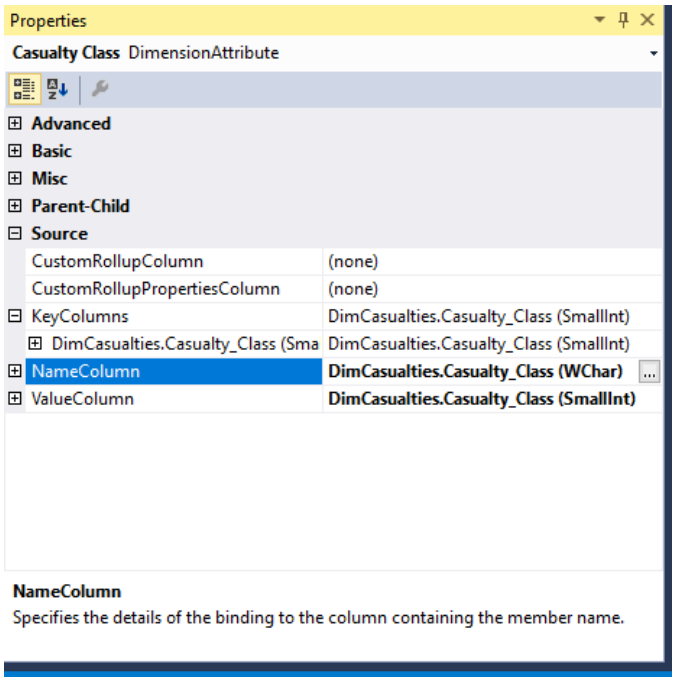

Properties
Year DimensionAttribute
Description
FormatString
ID
Year
Name
Year
Type
Years
Usage
Regular
Misc
Parent-Child
Source
CustomRollupColumn
(none)
CustomRollupPropertiesColumn
(none)
KeyColumns
(Collection)
DimDate.year (Integer)
DimDate.year (Integer)
DimDate.datetimekey (Double)
DimDate.datetimekey (Double)
NameColumn
DimDate.year (WChar)
ValueColumn
DimDate.year (Integer)
KeyColumns
Specifies the details of the binding to the column(s) containing the member key(s).

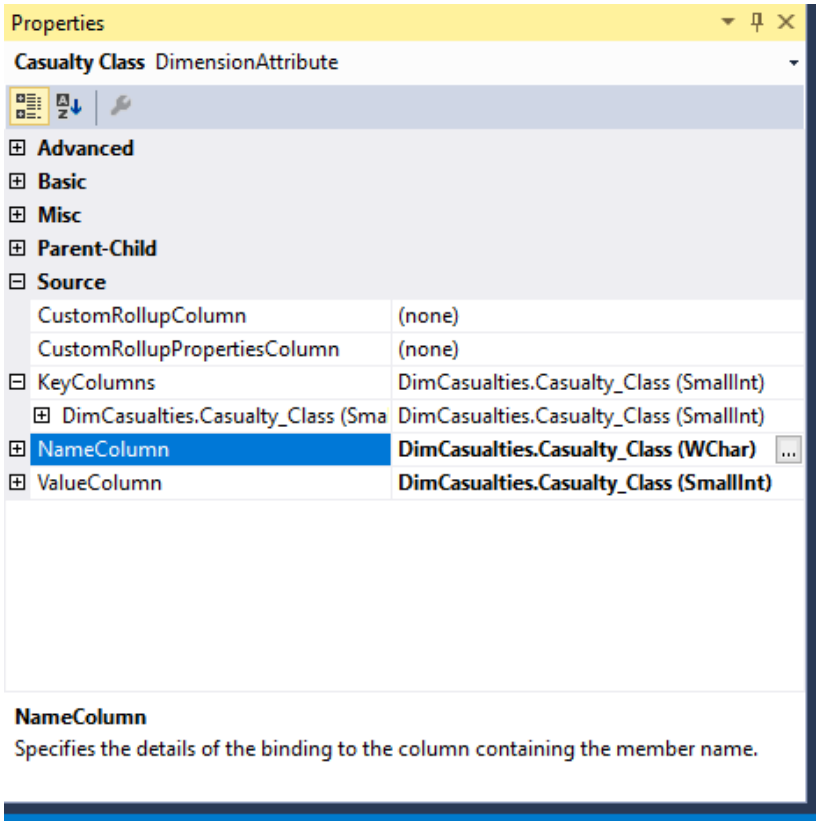
(ii) Casualties Dimension:

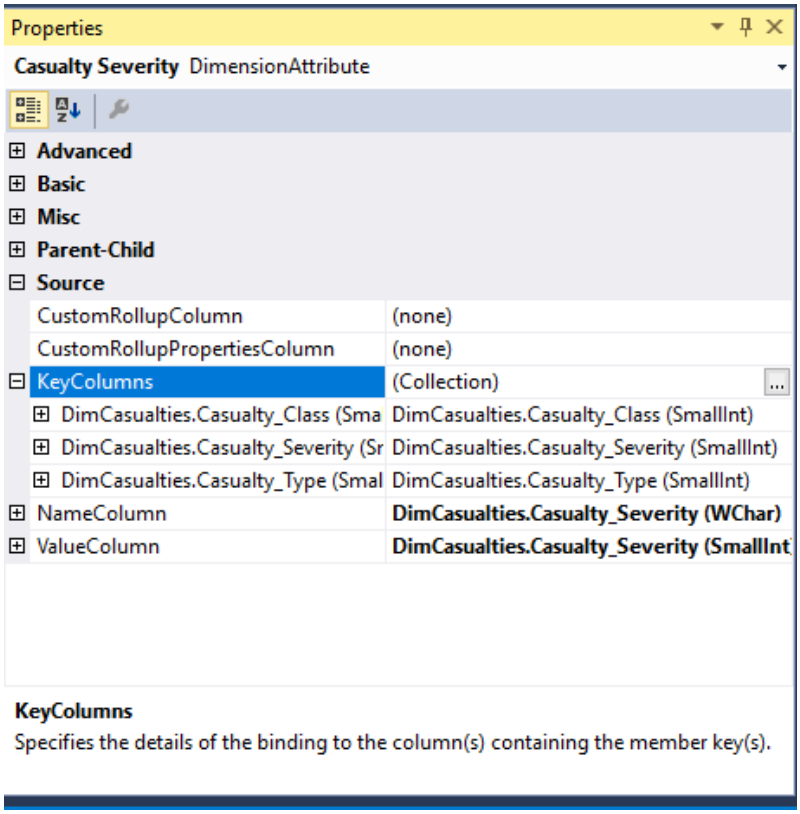
We created casualties hierarchies in casualties dimension



The properties of the attributes are as follows:




Properties
Casualty Class DimensionAttribute
Advanced
Basic
Misc
Parent-Child
Source
CustomRollupColumn
(none)
CustomRollupPropertiesColumn
(none)
KeyColumns
DimCasualties.Casualty_Class (SmallInt)
DimCasualties.Casualty_Class (Sma
DimCasualties.Casualty_Class (SmallInt)
NameColumn
DimCasualties.Casualty_Class (WChar)
ValueColumn
DimCasualties.Casualty_Class (SmallInt)
NameColumn
Specifies the details of the binding to the column containing the member name.


Properties
Casualty Severity DimensionAttribute
Advanced
Basic
Misc
Parent-Child
Source
CustomRollupColumn
(none)
CustomRollupPropertiesColumn
(none)
KeyColumns
(Collection)
DimCasualties.Casualty_Class (Sma
DimCasualties.Casualty_Class (SmallInt)
DimCasualties.Casualty_Severity (Sr
DimCasualties.Casualty_Severity (SmallInt)
DimCasualties.Casualty_Type (Smal
DimCasualties.Casualty_Type (SmallInt)
NameColumn
DimCasualties.Casualty_Severity (WChar)
ValueColumn
DimCasualties.Casualty_Severity (SmallInt
KeyColumns
Specifies the details of the binding to the column(s) containing the member key(s).

Step 8:Creating Partitions and Aggregations:

The main purpose of partitions and aggregations are to speedup the process of querying as we know data warehouse deals with large data and speeding up queries is vital.

We created partitions in Accident Master table based on years.

Since there are 11 years , 2005-2011. We decided to create 4 partitions containing 3 years each.

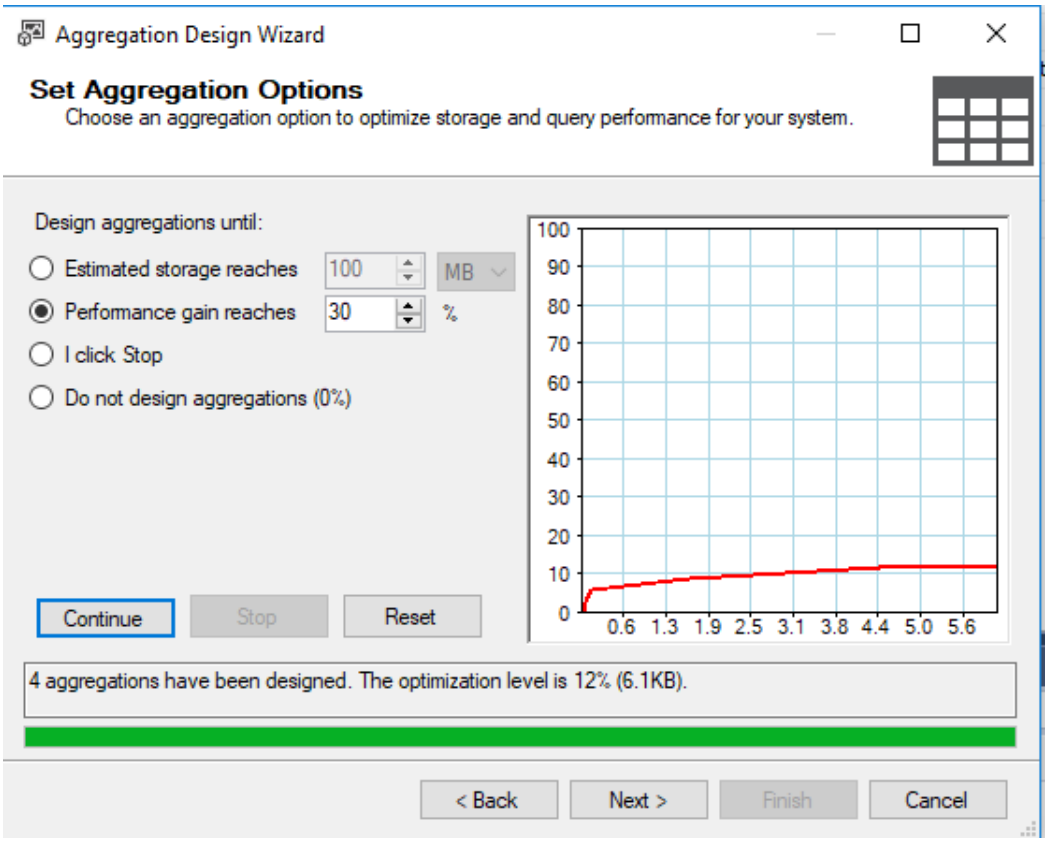
Accidents Master (4 Partitions)

	Partition Name	Source	Estimated Rows	Storage Mode	Aggregation Design
1	Accidents Master 2005 - 2007	SELECT [dbo].[Accidents_Master].[Accident_key],[dbo].[Accidents_Master].[A...	1506361	MOLAP	AggregationDesign - UK Accidents
2	Accidents Master 2008 - 2010	SELECT [dbo].[Accidents_Master].[Accident_key],[dbo].[Accidents_Master].[A...	0	MOLAP	AggregationDesign - UK Accidents
3	Accidents Master 2011- 2013	SELECT [dbo].[Accidents_Master].[Accident_key],[dbo].[Accidents_Master].[A...	0	MOLAP	AggregationDesign - UK Accidents
4	Accidents Master greater than 2013	SELECT [dbo].[Accidents_Master].[Accident_key],[dbo].[Accidents_Master].[A...	0	MOLAP	AggregationDesign - UK Accidents

New Partition... Storage Settings...

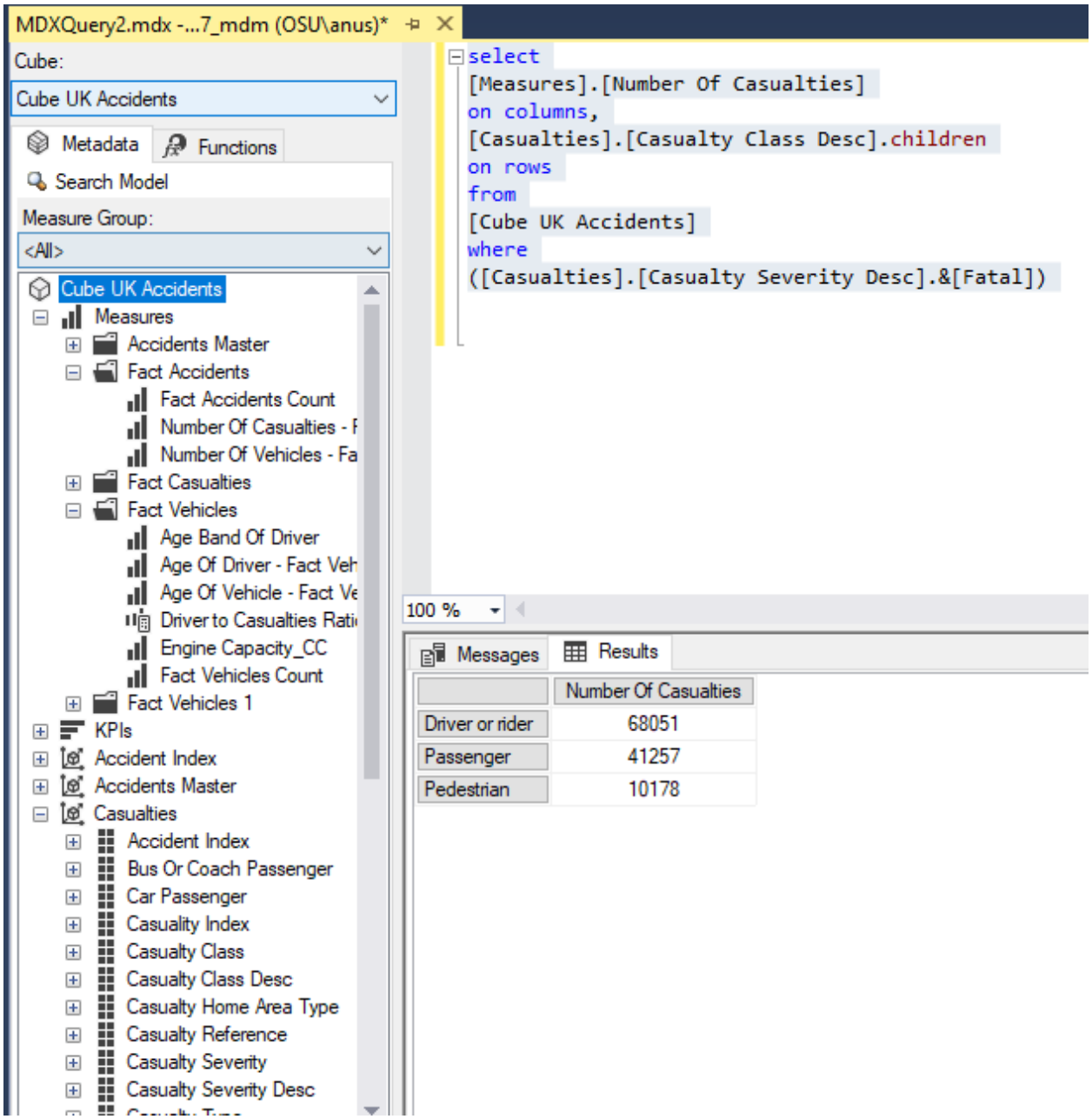
We decided to do 30% aggregation as shown below,

	Aggregations	Estimated Partition Size	Partitions
Accidents Master (1 Aggregation Design)			
AggregationDesign - UK Accidents	4	1506361	Accidents Master 2005 - 2007, Accidents Master 2008 - 2010, Accidents Master 2011- 2013, ...



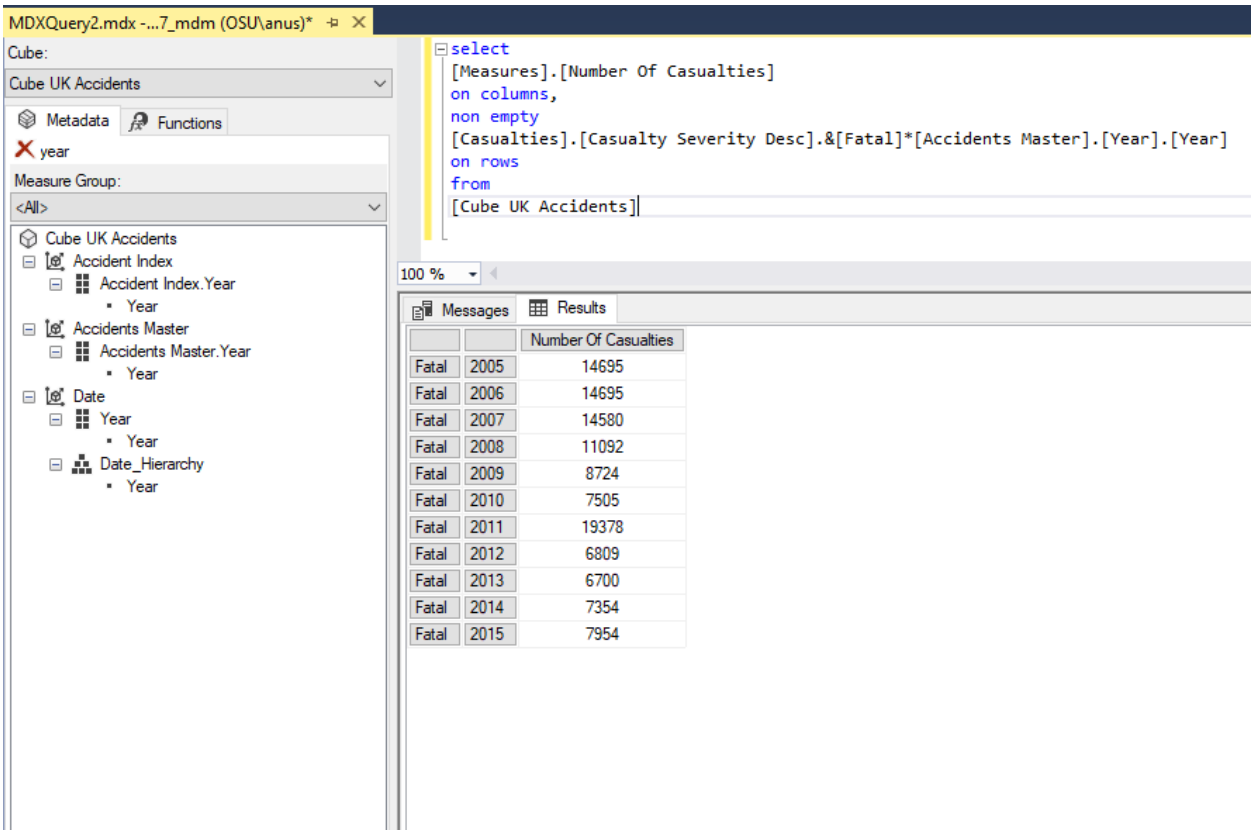
Step 9:Designing and Executing MDX Queries:

1. What are the total number of casualties in each class?



Driver or Rider are the most affected in casualties than other classes .

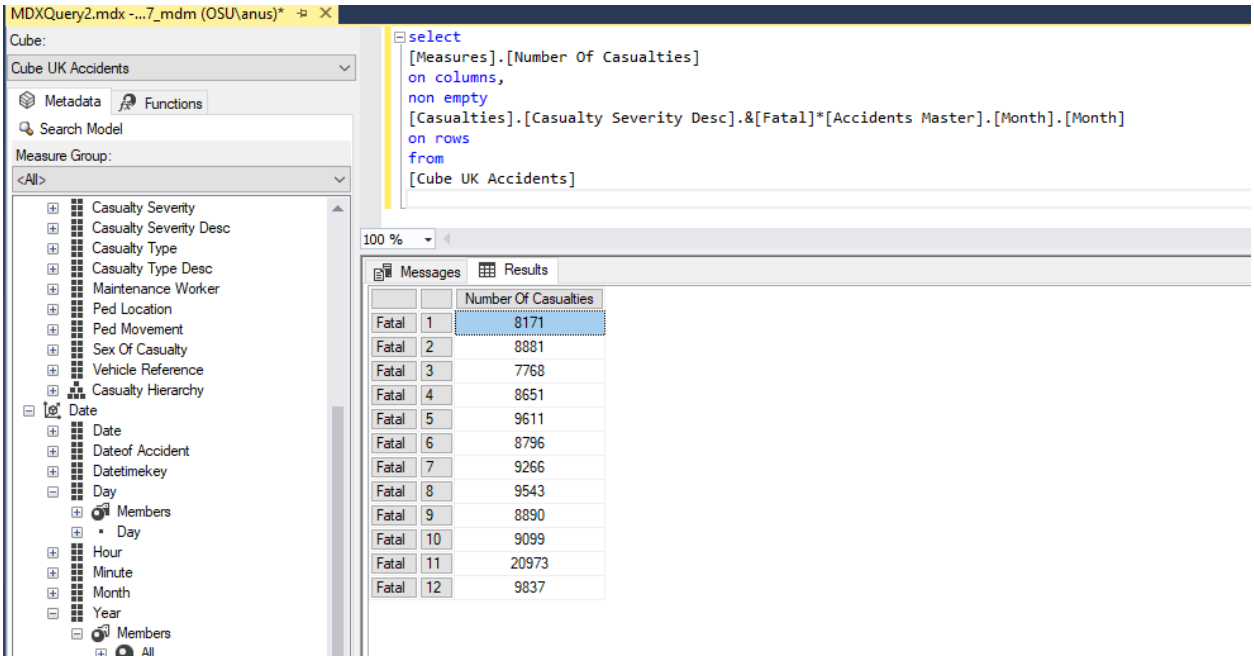
2. Number of fatal accidents in each year?



		Number Of Casualties
Fatal	2005	14695
Fatal	2006	14695
Fatal	2007	14580
Fatal	2008	11092
Fatal	2009	8724
Fatal	2010	7505
Fatal	2011	19378
Fatal	2012	6809
Fatal	2013	6700
Fatal	2014	7354
Fatal	2015	7954

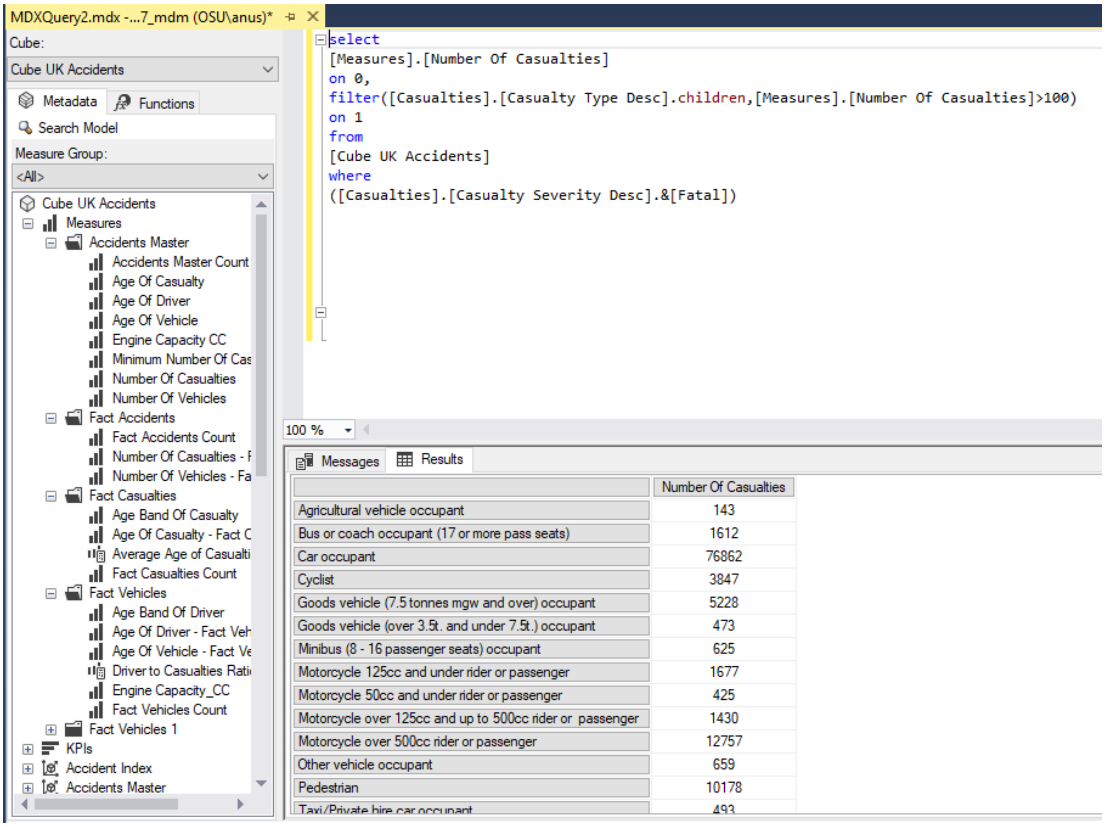
From the query we see that, 2005 and 2006 have the most number of casualties.

3. Total number of casualties in each month of all years



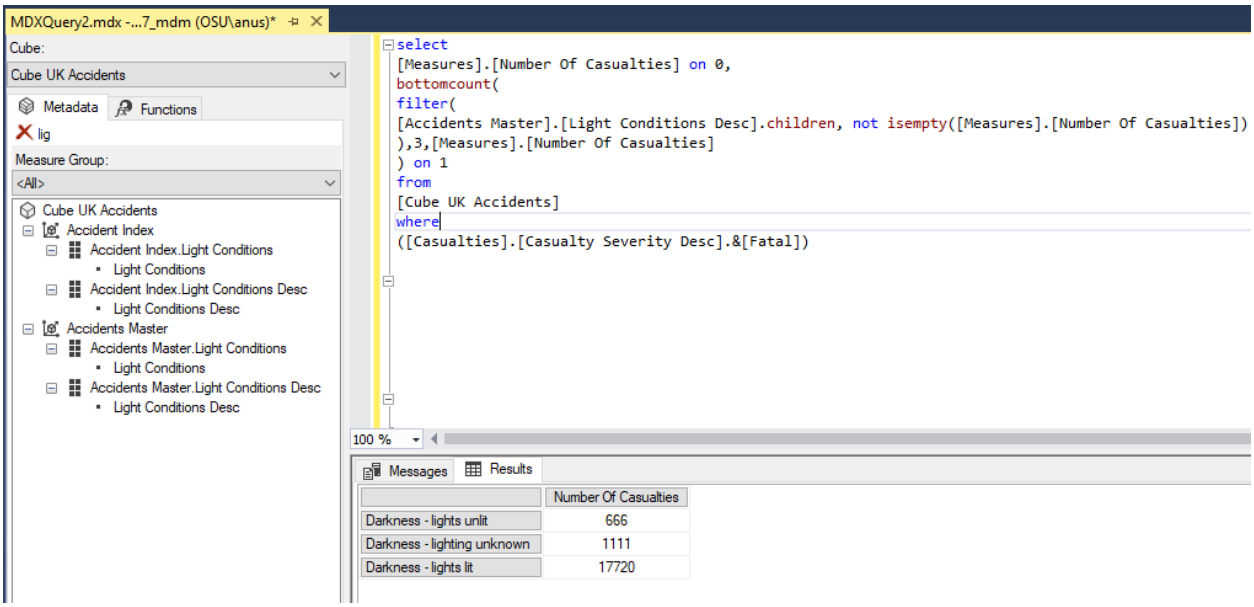
We see that most casualties happen during November, Weather might have played an important role in these accidents.

4. Which vehicles involved in most of the casualties:



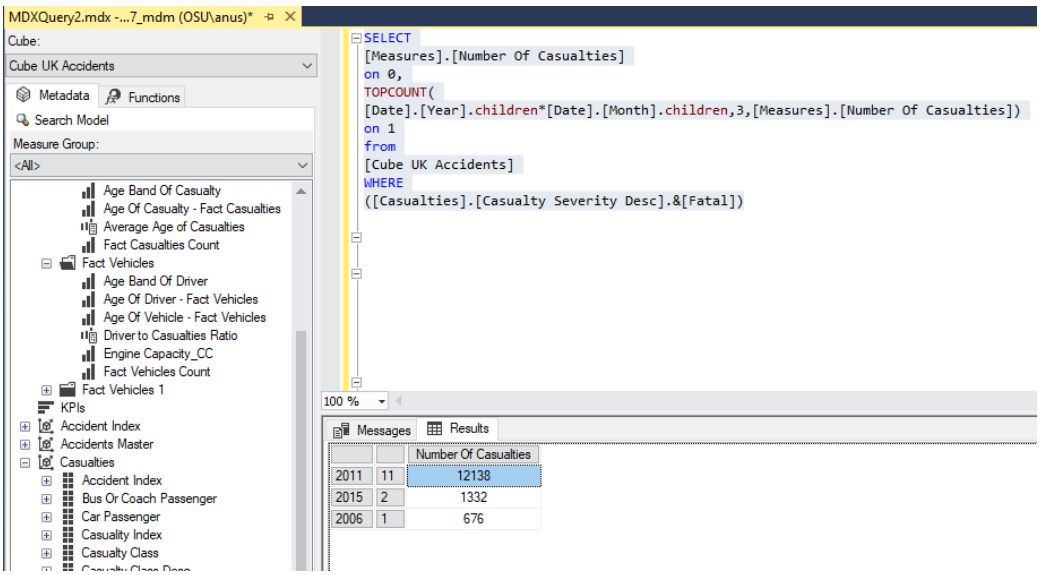
We see that car occupants are the most affected by casualties

5. Which light conditions have the most casualties?



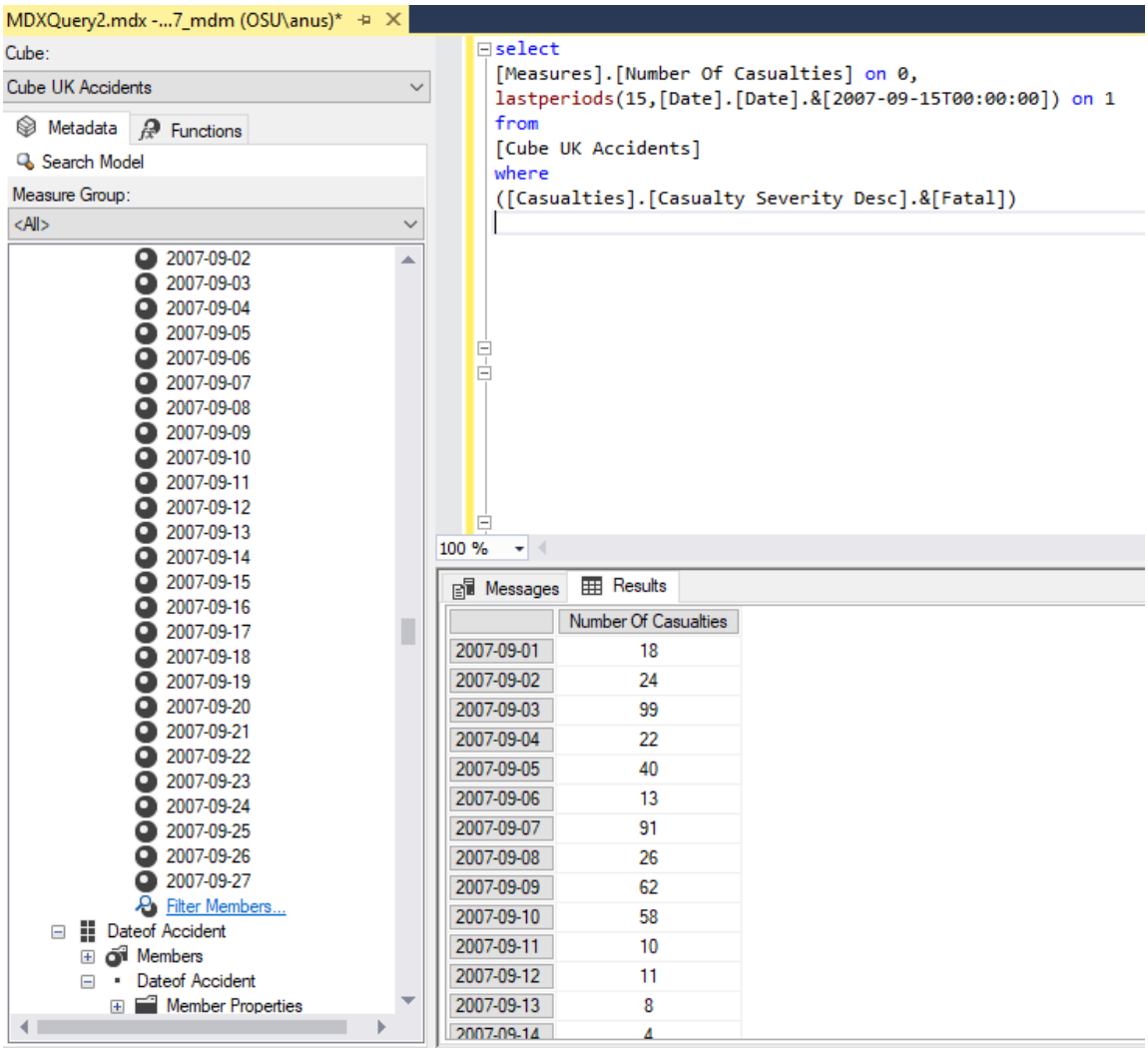
We can understand from the query that dark roads with lights lit suffer the most number of casualties.

6. What are the top 3 months in which the casualties occur?

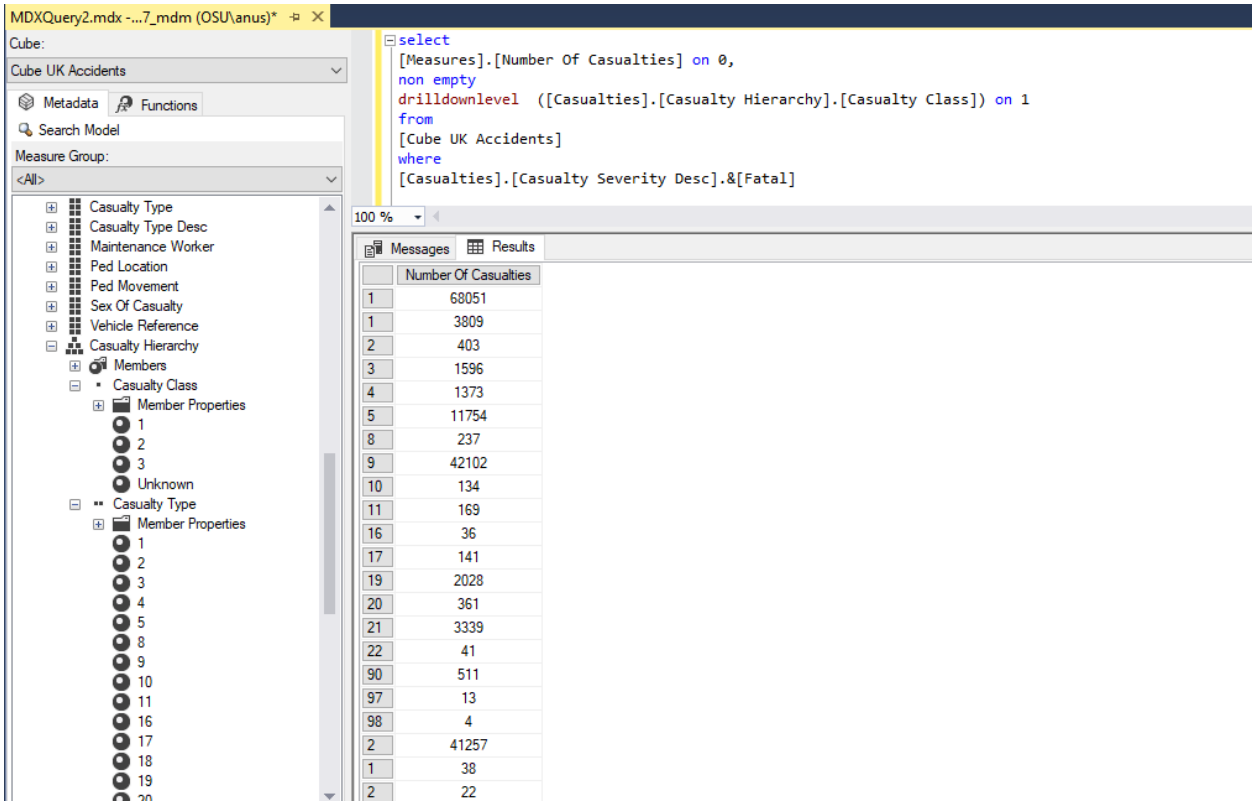


Casualties occur the most in November, followed by February and January.

7. What are the last 15 periods before September 15, 2007 that the casualties occurred?

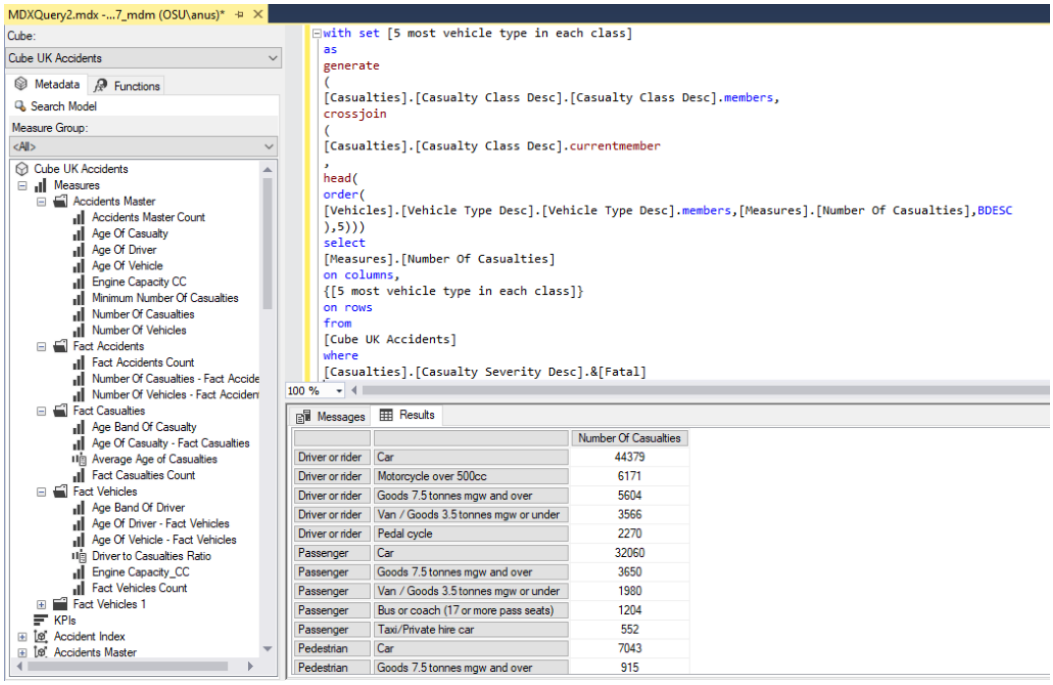


8. Drilldown the casualties hierarchy where the casualty severity is fatal



We see that in level one , the descendants are from 1 to 98 followed by the hierarchy level 2 with number of fatalities.

9. What are the top 5 vehicles in each types that suffer fatal accidents



10. Rank each vehicles according to the number of fatalities

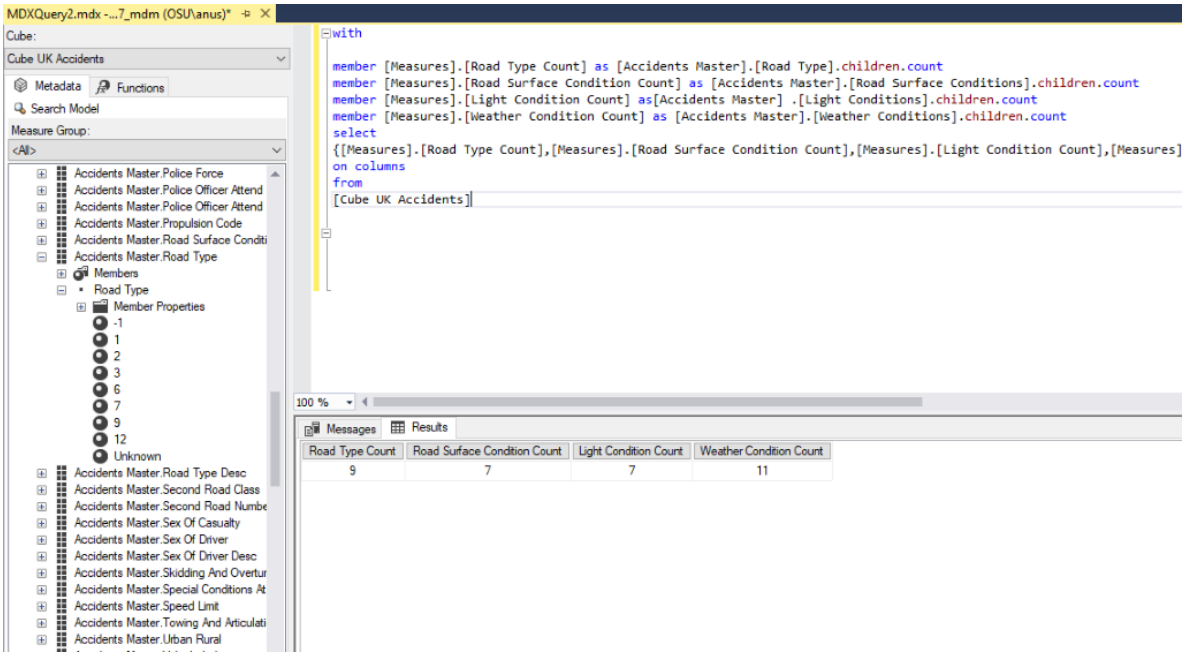
```
with
set [Top 5 Vehicle Types] as
topcount(
[Vehicles].[Vehicle Type Desc].children,5,[Measures].[Number Of Casualties]
)
member [Measures].[Months Rank] as
rank([Vehicles].[Vehicle Type Desc].currentmember,[Top 5 Vehicle Types])
select
{[Measures].[Number Of Casualties],[Measures].[Months Rank]}
on columns,
[Top 5 Vehicle Types]
on rows
from
[Cube UK Accidents]
where
([Casualties].[Casualty Severity Desc].&[Fatal])
```

100 %

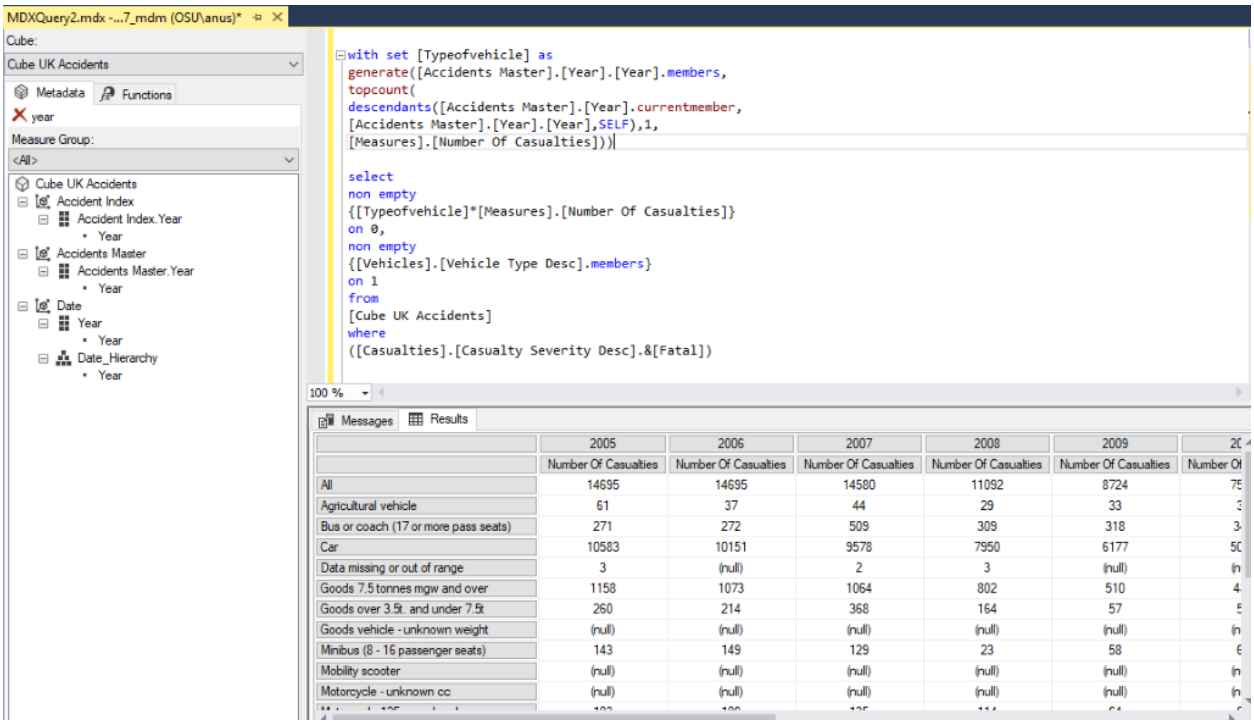
Messages | Results

	Number Of Casualties	Months Rank
Car	83482	1
Goods 7.5 tonnes mgw and over	10169	2
Motorcycle over 500cc	6932	3
Van / Goods 3.5 tonnes mgw or under	6174	4
Bus or coach (17 or more pass seats)	3048	5

11. Display the count of Road type, Road surface condition, Light condition, Weather condition to know the types



12. Number of fatalities that occur each year in each vehicle



Data Mining:

As we know that data mining is process of knowledge discovery from the large databases and datasets that can be used for further analysis in intelligent decision making. Without proper data mining and knowledge discovery, there is no purpose of maintaining large databases and Enterprise warehouses with large setting and operating costs. Data mining focuses on specific machine learning methods and procedures to empower predictive modeling and pattern finding hence enabling intelligent decision making.

With SQL Server Management Studio, the mining models can be created using Microsoft specific language called Data Mining Extensions abbreviated as (DMX) which will be used for creating mining structures, mining models, to train, test and evaluate the built models.

In the project, we have used UK Accidents Database. The predictive models will be built to predict the Accident severity. This will be predicted based on various independent variables which will be discussed in later part.

Since the predictor variable is Accident Severity, we have chosen to use

Decision Tree: As it performs better for both classification and regression problems

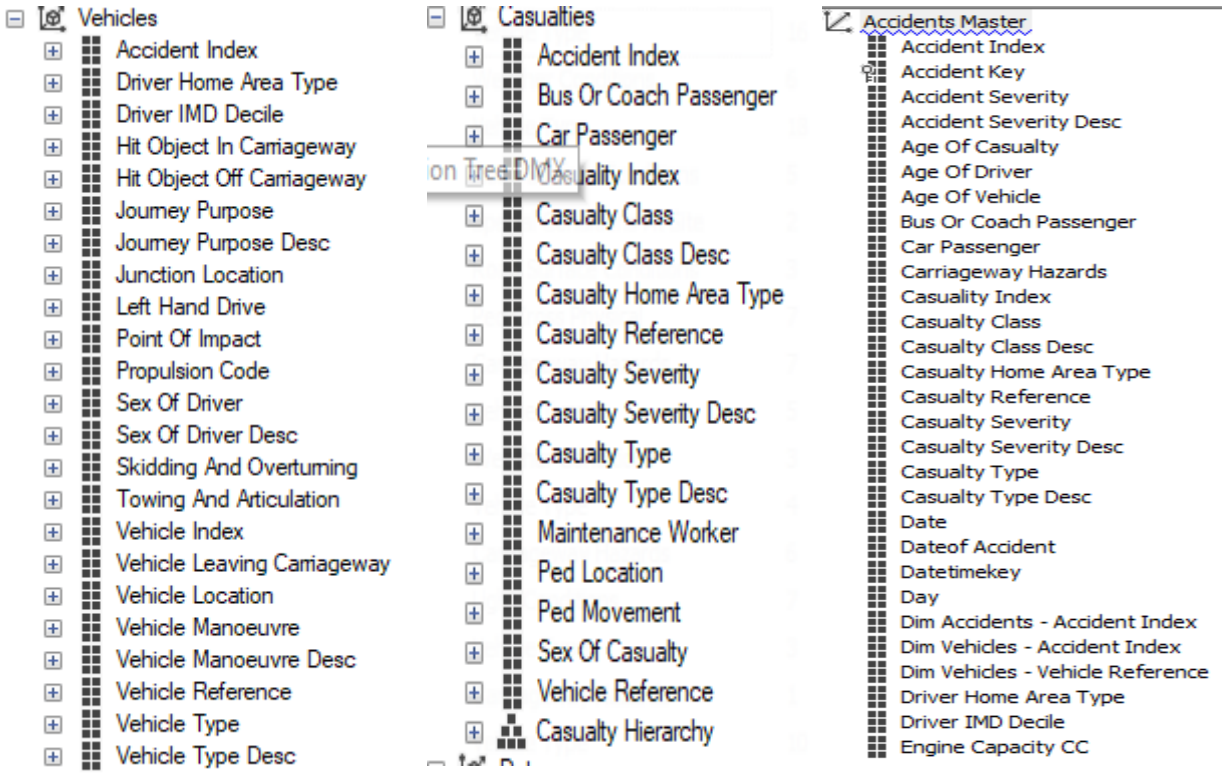
Logistic Regression: As logistic regression is the most sought classification problem for categorical variables

Neural Networks: We chose Neural Networks as we believe Neural Networks to be better for both supervised and unsupervised learning in classification problems.

Predictor and Target variables:

As we have chosen the Target variable as accident severity, we thought it would be interesting to know which type of accident lead to serious injury.

We built the model using SSMS, considering the following tables,



As using all attributes can lead to underperformance and overfitting of the model. We chose to use several variables that were considered helpful in predicting the accident severity.

```
\
[Accident Index],
[Accident Severity] PREDICT,
[Carriageway Hazards],
[First Road Class],
[Junction Control],
[Junction Detail],
[Light Conditions],
[Ped Cross Physical],
[Road Surface Conditions],
[Road Type],
[Special Conditions At Site],
[Speed Limit],
[Urban Rural],
[Weather Conditions],
[Sex of Casualty],
[Journey Purpose],
[Vehicle Type]
```

We chose the above variables to predict the accident severity.

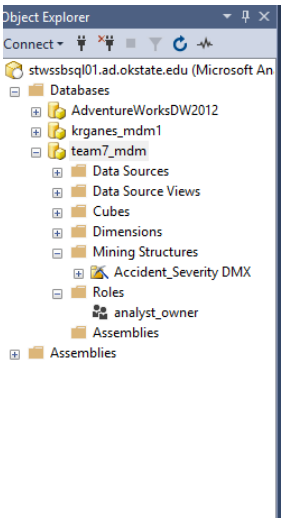
Building Mining Models:

Creating Mining Structure:

We started with creating the Accident-Severity DMX

```
CREATE MINING STRUCTURE [Accident_Severity 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 Physical] TEXT DISCRETE,
[Road Surface Conditions] TEXT DISCRETE,
[Road Type] TEXT DISCRETE,
[Special Conditions At Site] TEXT DISCRETE,
[Speed Limit] TEXT DISCRETE,
[Urban Rural] TEXT DISCRETE,
[Weather Conditions] TEXT DISCRETE,
[Sex of Casualty] TEXT DISCRETE,
[Journey Purpose] TEXT DISCRETE,
[Vehicle Type] TEXT DISCRETE
)
WITH HOLDOUT (30 PERCENT)
```

The created mining structure is shown below,



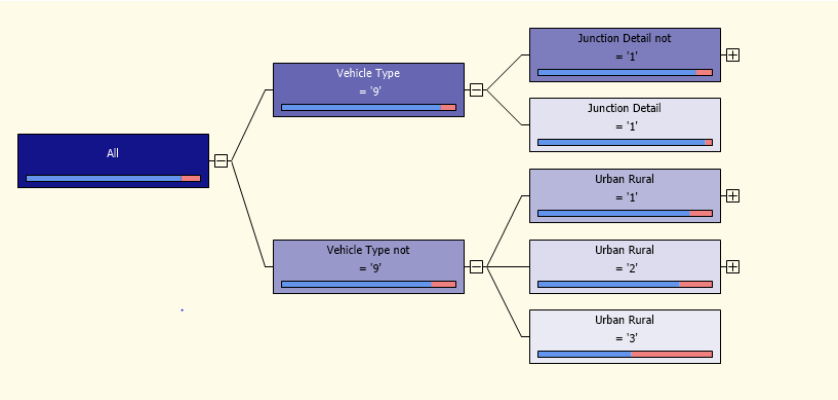
Building Models:

Decision Trees:

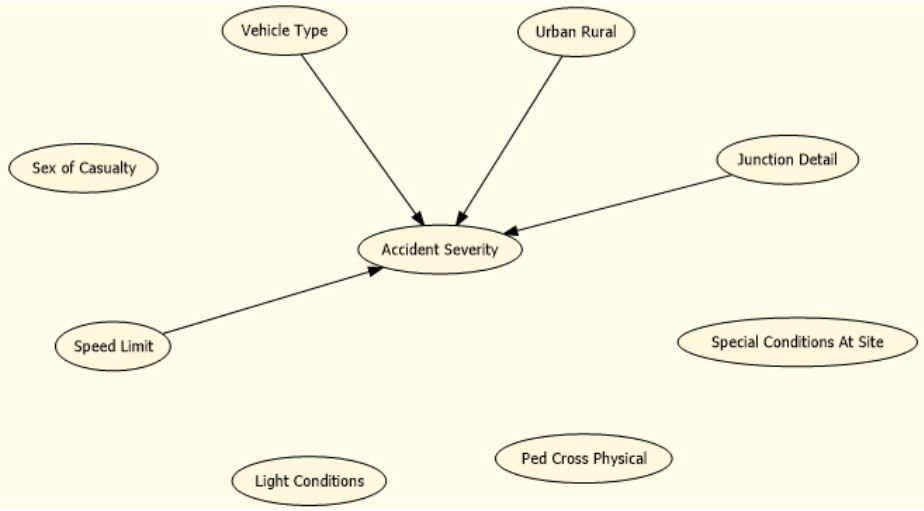
```
ALTER MINING STRUCTURE [Accidents DMX]
ADD MINING MODEL [Decision Tree DMX]
( [Accident Index],
[Accident Severity] PREDICT,
[Carriageway Hazards],
[First Road Class],
[Junction Control],
[Junction Detail],
[Light Conditions],
[Ped Cross Physical],
[Road Surface Conditions],
[Road Type],
[Special Conditions At Site],
[Speed Limit],
[Urban Rural],
[Weather Conditions],
[Sex of Casualty],
[Journey Purpose],
[Vehicle Type] )
USING Microsoft_Decision_Trees
```

```
INSERT INTO MINING STRUCTURE [Accidents DMX]
( [Accident Index],  [Accident Severity],
[Carriageway Hazards], [First Road Class],
[Junction Control],  [Junction Detail],
[Light Conditions],  [Ped Cross Physical],
[Road Surface Conditions],
[Road Type],  [Special Conditions At Site],
[Speed Limit],  [Urban Rural],
[Weather Conditions], [Sex of Casualty],
[Journey Purpose], [Vehicle Type] )

OPENQUERY([UK Accidents Database],
'SELECT TOP 100000 DimAccidents.Accident_Index,
Accident_Severity,
Carriageway_Hazards,
First_Road_Class,
Junction_Control,
Junction_Detail,
Light_Conditions,
Ped_Cross_Physical,
Road_Surface_Conditions,
Road_Type,
Special_Conditions_At_Site,
Speed_Limit,  Urban_Rural,
Weather_Conditions, Sex_of_Casualty,
Journey_Purpose, Vehicle_Type FROM DimAccidents,
DimCasualties, DimVehicles
where  DimAccidents.Accident_Index = DimCasualties.Accident_Index AND
DimAccidents.Accident_Index = DimVehicles.Accident_Index AND
Accident_Severity != 1 AND
Carriageway_Hazards != -1 AND
First_Road_Class != 6 AND
Junction_Control != -1 AND
Junction_Detail != -1 AND
Light_Conditions != -1 AND
Ped_Cross_Physical != -1 AND
Road_Surface_Conditions != -1 AND
Road_Type != -1 AND
Road_Type != 9 AND
Special_Conditions_At_Site != 9 AND
Weather_Conditions != -1 AND
Sex_of_Casualty != -1 AND
Vehicle_Type != -1 AND
Journey_Purpose != 15 AND
Journey_Purpose != -1 AND
Journey_Purpose != 6 ')
```

Results :



We see from the decision the results that vehicle type= 9 which is car impacts the severity of the accident. If a person, who is a male and is traveling in car with a speed greater than 60 is prone to severe injury.



The important model variables from decision tree determining Accident Severity are Speed limit, vehicle type, urban rural and junction detail.

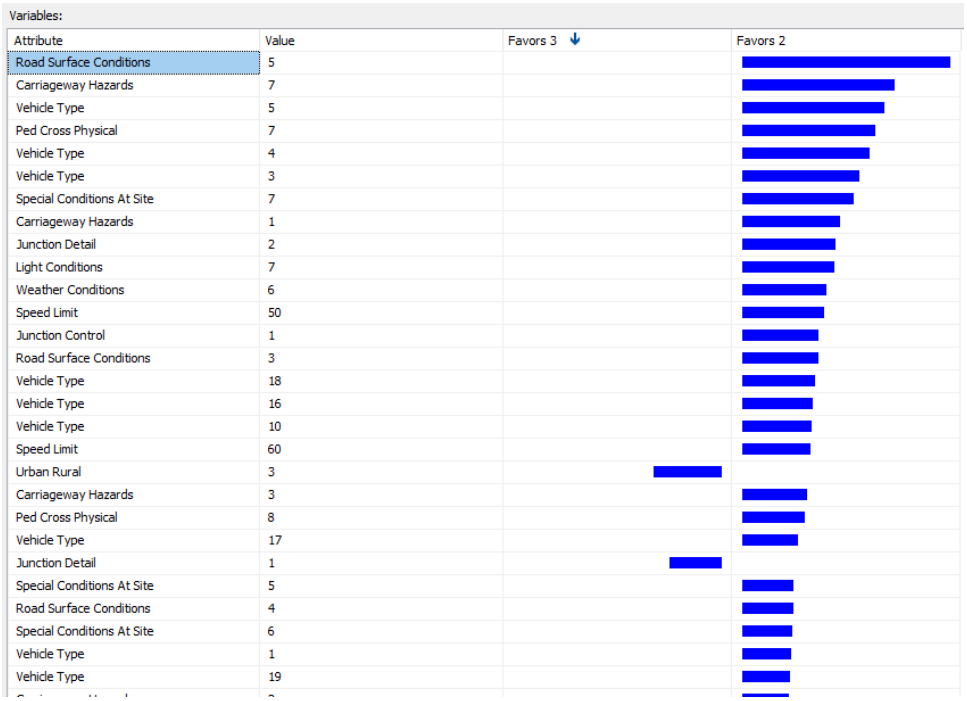
Neural Networks:

```
//Neural Networks Model

ALTER MINING STRUCTURE [Accidents DMX]
ADD MINING MODEL [Neural Networks DMX]
( [Accident Index],
[Accident Severity] PREDICT,
[Carriageway Hazards],
[First Road Class],
[Junction Control],
[Junction Detail],
[Light Conditions],
[Ped Cross Physical],
[Road Surface Conditions],
[Road Type],
[Special Conditions At Site],
[Speed Limit],
[Urban Rural],
[Weather Conditions],
[Sex of Casualty],
[Journey Purpose],
[Vehicle Type] )
USING Microsoft_Neural_Network
```

```
INSERT INTO MINING STRUCTURE [Accidents DMX]
( [Accident Index],  [Accident Severity],
[Carriageway Hazards], [First Road Class],
[Junction Control],  [Junction Detail],
[Light Conditions],  [Ped Cross Physical],
[Road Surface Conditions],
[Road Type],  [Special Conditions At Site],
[Speed Limit],  [Urban Rural],
[Weather Conditions], [Sex of Casualty],
[Journey Purpose], [Vehicle Type] )

OPENQUERY([UK Accidents Database],
'SELECT TOP 100000 DimAccidents.Accident_Index,
Accident_Severity,
Carriageway_Hazards,
First_Road_Class,
Junction_Control,
Junction_Detail,
Light_Conditions,
Ped_Cross_Physical,
Road_Surface_Conditions,
Road_Type,
Special_Conditions_At_Site,
Speed_Limit,  Urban_Rural,
Weather_Conditions, Sex_of_Casualty,
Journey_Purpose, Vehicle_Type FROM DimAccidents,
DimCasualties, DimVehicles
where  DimAccidents.Accident_Index = DimCasualties.Accident_Index AND
DimAccidents.Accident_Index = DimVehicles.Accident_Index AND
Accident_Severity != 1 AND
Carriageway_Hazards != -1 AND
First_Road_Class != 6 AND
Junction_Control != -1 AND
Junction_Detail != -1 AND
Light_Conditions != -1 AND
Ped_Cross_Physical != -1 AND
Road_Surface_Conditions != -1 AND
Road_Type != -1 AND
Road_Type != 9 AND
Special_Conditions_At_Site != 9 AND
Weather_Conditions != -1 AND
Sex_of_Casualty != -1 AND
Vehicle_Type != -1 AND
Journey_Purpose != 15 AND
Journey_Purpose != -1 AND
Journey_Purpose != 6 ')
```

Results:

The results of Neural Networks are as shown below,



Variables:

Attribute	Value	Favors 3	Favors 2
Road Surface Conditions	5		
Carriageway Hazards	7		
Vehicle Type	5		
Ped Cross Physical	7		
Vehicle Type	4		
Vehicle Type	3		
Special Conditions At Site	7		
Carriageway Hazards	1		
Junction Detail	2		
Light Conditions	7		
Weather Conditions	6		
Speed Limit	50		
Junction Control	1		
Road Surface Conditions	3		
Vehicle Type	18		
Vehicle Type	16		
Vehicle Type	10		
Speed Limit	60		
Urban Rural	3		
Carriageway Hazards	3		
Ped Cross Physical	8		
Vehicle Type	17		
Junction Detail	1		
Special Conditions At Site	5		
Road Surface Conditions	4		
Special Conditions At Site	6		
Vehicle Type	1		
Vehicle Type	19		

We can see that Road surface conditions being , followed by vehicle load on the carriage way and vehicle type stand as most important factors in determining the severity of the accident being serious.

Logistic Regression:

```
//Logistic Regression Model

ALTER MINING STRUCTURE [Accidents DMX]
ADD MINING MODEL [Logistic Regression DMX]
( [Accident Index],
[Accident Severity] PREDICT,
[Carriageway Hazards],
[First Road Class],
[Junction Control],
[Junction Detail],
[Light Conditions],
[Ped Cross Physical],
[Road Surface Conditions],
[Road Type],
[Special Conditions At Site],
[Speed Limit],
[Urban Rural],
[Weather Conditions],
[Sex of Casualty],
[Journey Purpose],
[Vehicle Type] )
USING Microsoft_Logistic_Regression
```

```
INSERT INTO MINING STRUCTURE [Accidents DMX]
( [Accident Index],  [Accident Severity],
[Carriageway Hazards], [First Road Class],
[Junction Control],  [Junction Detail],
[Light Conditions],  [Ped Cross Physical],
[Road Surface Conditions],
[Road Type],  [Special Conditions At Site],
[Speed Limit],  [Urban Rural],
[Weather Conditions], [Sex of Casualty],
[Journey Purpose], [Vehicle Type] )

OPENQUERY([UK Accidents Database],
'SELECT TOP 100000 DimAccidents.Accident_Index,
Accident_Severity,
Carriageway_Hazards,
First_Road_Class,
Junction_Control,
Junction_Detail,
Light_Conditions,
Ped_Cross_Physical,
Road_Surface_Conditions,
Road_Type,
Special_Conditions_At_Site,
Speed_Limit,  Urban_Rural,
Weather_Conditions, Sex_of_Casualty,
Journey_Purpose, Vehicle_Type FROM DimAccidents,
DimCasualties, DimVehicles
where  DimAccidents.Accident_Index = DimCasualties.Accident_Index AND
DimAccidents.Accident_Index = DimVehicles.Accident_Index AND
Accident_Severity != 1 AND
Carriageway_Hazards != -1 AND
First_Road_Class != 6 AND
Junction_Control != -1 AND
Junction_Detail != -1 AND
Light_Conditions != -1 AND
Ped_Cross_Physical != -1 AND
Road_Surface_Conditions != -1 AND
Road_Type != -1 AND
Road_Type != 9 AND
Special_Conditions_At_Site != 9 AND
Weather_Conditions != -1 AND
Sex_of_Casualty != -1 AND
Vehicle_Type != -1 AND
Journey_Purpose != 15 AND
Journey_Purpose != -1 AND
Journey_Purpose != 6 ')
```

Results:

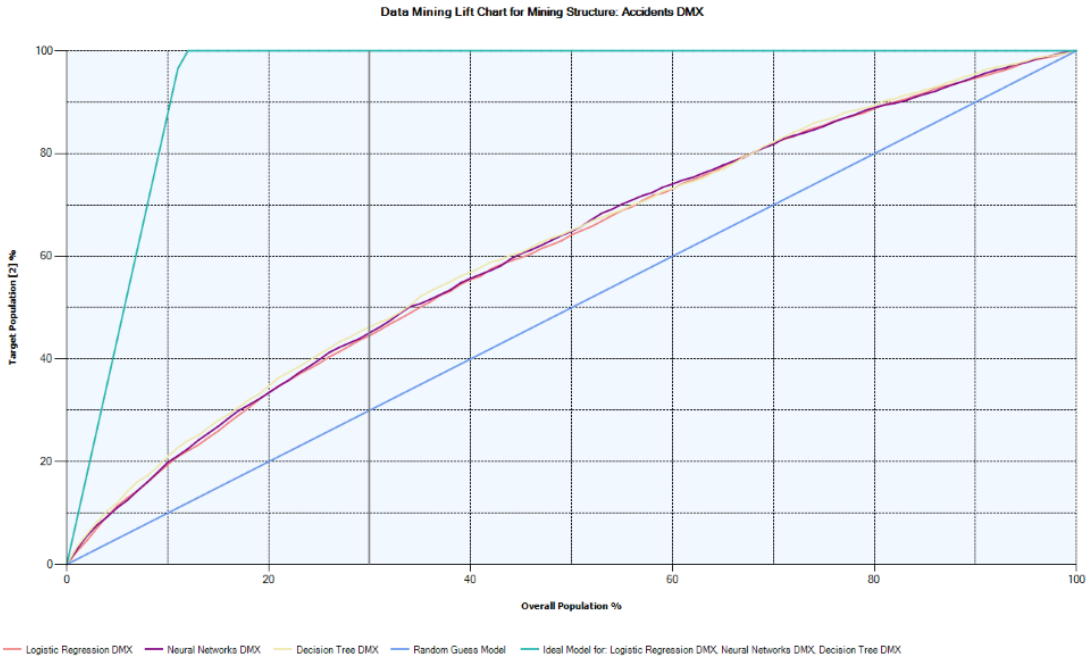
Variables:

Attribute	Value	Favors 3	Favors 2
Vehicle Type	16		
Weather Conditions	6		
Vehicle Type	18		
Road Surface Conditions	5		
Special Conditions At Site	2		
Road Surface Conditions	3		
Ped Cross Physical	7		
Carriageway Hazards	7		
Vehicle Type	5		
Weather Conditions	3		
Vehicle Type	4		
Carriageway Hazards	6		
Light Conditions	7		
Vehicle Type	3		
Carriageway Hazards	1		
Vehicle Type	10		
Junction Control	1		
Junction Detail	2		
Special Conditions At Site	4		
Special Conditions At Site	3		
Light Conditions	5		
Urban Rural	3		
Special Conditions At Site	5		
Weather Conditions	9		
Speed Limit	50		

We see from the screenshot above that vehicle type and surface conditions play an important role in accident severity to be fatal . Weather conditions, road surface conditions play important role in determining the accident to be serious.

Comparison of Results:

Lift Chart:



Mining Legend

Population percentage: 30.00%

Series, Model	Score	Target population	Predict probability
Logistic Regression DMX	0.64	44.55%	13.14%
Neural Networks DMX	0.65	45.10%	12.33%
Decision Tree DMX	0.65	46.21%	12.96%
Random Guess Model		30.00%	
Ideal Model for: Logistic Regression D...		100.00%	

The lift score of all the models are nearly the same. We chose the population percentage of 30% .

This can be interpreted as follows:

The logistic model will correctly identify 44.55% of the people who are seriously injured in the entire population. To identify the severe injury among the people , we would use query to retrieve cases with a predict probability of 13.14%

The Neural Network will correctly identify 45.10% of the people who are seriously injured in the entire population. To identify the severe injury among the people , we would use query to retrieve cases with a predict probability of 12.33%

The Decision Tree will correctly identify 46.21% of the people who are seriously injured in the entire population. To identify the severe injury among the people , we would use query to retrieve cases with a predict probability of 12.96%

From the score, the performance of Neural Network and Decision Tree are better but from the prediction probability it can be concluded that Decision Tree performs the best.

Classification Matrix:

Counts for Logistic Regression DMX on Accident Severity			
	Predicted	3 (Actual)	2 (Actual)
	3	26534	3405
	2	47	14
Counts for Neural Networks DMX on Accident Severity			
	Predicted	3 (Actual)	2 (Actual)
	3	26579	3419
	2	2	0
Counts for Decision Tree DMX on Accident Severity			
	Predicted	3 (Actual)	2 (Actual)
	3	26575	3373
	2	6	46

S.No	Model Name	Correct Classification Accuracy
1	Decision Tree	0.887
2	Logistic Regression	0.884
3	Neural Networks	0.885

Classification model proves that Decision Tree has better accuracy.

Cross Validation:

A 10 fold cross validation on injury severity=2 with a threshold of 0.3 was performed.

Logistic Regression DMX				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	6999	Classification	True Positive	26
2	7000	Classification	True Positive	37
3	7001	Classification	True Positive	35
4	7000	Classification	True Positive	46
5	7000	Classification	True Positive	42
6	7000	Classification	True Positive	47
7	7000	Classification	True Positive	35
8	7000	Classification	True Positive	50
9	7000	Classification	True Positive	41
10	7000	Classification	True Positive	39
			Average Standard Deviation	39.8001 6.6751
1	6999	Classification	False Positive	82
2	7000	Classification	False Positive	99
3	7001	Classification	False Positive	61
4	7000	Classification	False Positive	94
5	7000	Classification	False Positive	77
6	7000	Classification	False Positive	101
7	7000	Classification	False Positive	86
8	7000	Classification	False Positive	104
9	7000	Classification	False Positive	98
10	7000	Classification	False Positive	115
			Average Standard Deviation	91.6997 14.7113
1	6999	Classification	True Negative	6112
2	7000	Classification	True Negative	6096
3	7001	Classification	True Negative	6134
4	7000	Classification	True Negative	6101
5	7000	Classification	True Negative	6118
6	7000	Classification	True Negative	6094
7	7000	Classification	True Negative	6109
8	7000	Classification	True Negative	6091
9	7000	Classification	True Negative	6097
10	7000	Classification	True Negative	6080

			Average	6103.2003
			Standard Deviation	14.6483
1	6999	Classification	False Negative	779
2	7000	Classification	False Negative	768
3	7001	Classification	False Negative	771
4	7000	Classification	False Negative	759
5	7000	Classification	False Negative	763
6	7000	Classification	False Negative	758
7	7000	Classification	False Negative	770
8	7000	Classification	False Negative	755
9	7000	Classification	False Negative	764
10	7000	Classification	False Negative	766
			Average	765.2999
			Standard Deviation	6.7534
1	6999	Likelihood	Log Score	-0.3476
2	7000	Likelihood	Log Score	-0.3485
3	7001	Likelihood	Log Score	-0.3469
4	7000	Likelihood	Log Score	-0.3453
5	7000	Likelihood	Log Score	-0.3472
6	7000	Likelihood	Log Score	-0.3508
7	7000	Likelihood	Log Score	-0.349
8	7000	Likelihood	Log Score	-0.3461
9	7000	Likelihood	Log Score	-0.3499
10	7000	Likelihood	Log Score	-0.3485
			Average	-0.348
			Standard Deviation	0.0016
1	6999	Likelihood	Lift	0.0093
2	7000	Likelihood	Lift	0.0083
3	7001	Likelihood	Lift	0.0102
4	7000	Likelihood	Lift	0.0116
5	7000	Likelihood	Lift	0.0097
6	7000	Likelihood	Lift	0.006
7	7000	Likelihood	Lift	0.0079
8	7000	Likelihood	Lift	0.0107
9	7000	Likelihood	Lift	0.0069
10	7000	Likelihood	Lift	0.0083
			Average	0.0089
			Standard Deviation	0.0016

1	6999	Likelihood	Root Mean Square Error	0.1418
2	7000	Likelihood	Root Mean Square Error	0.1448
3	7001	Likelihood	Root Mean Square Error	0.1423
4	7000	Likelihood	Root Mean Square Error	0.148
5	7000	Likelihood	Root Mean Square Error	0.1419
6	7000	Likelihood	Root Mean Square Error	0.1462
7	7000	Likelihood	Root Mean Square Error	0.1414
8	7000	Likelihood	Root Mean Square Error	0.1461
9	7000	Likelihood	Root Mean Square Error	0.1471
10	7000	Likelihood	Root Mean Square Error	0.149
			Average	0.1449
			Standard Deviation	0.0027
Neural Networks DMX				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	6999	Classification	True Positive	22
2	7000	Classification	True Positive	25
3	7001	Classification	True Positive	15
4	7000	Classification	True Positive	42
5	7000	Classification	True Positive	39
6	7000	Classification	True Positive	34
7	7000	Classification	True Positive	20
8	7000	Classification	True Positive	34
9	7000	Classification	True Positive	33
10	7000	Classification	True Positive	33
			Average	29.6999
			Standard Deviation	8.2953
1	6999	Classification	False Positive	71
2	7000	Classification	False Positive	74
3	7001	Classification	False Positive	38
4	7000	Classification	False Positive	95

5	7000	Classification	False Positive	64
6	7000	Classification	False Positive	68
7	7000	Classification	False Positive	47
8	7000	Classification	False Positive	51
9	7000	Classification	False Positive	75
10	7000	Classification	False Positive	97
			Average Standard Deviation	67.9995 18.1938
1	6999	Classification	True Negative	6123
2	7000	Classification	True Negative	6121
3	7001	Classification	True Negative	6157
4	7000	Classification	True Negative	6100
5	7000	Classification	True Negative	6131
6	7000	Classification	True Negative	6127
7	7000	Classification	True Negative	6148
8	7000	Classification	True Negative	6144
9	7000	Classification	True Negative	6120
10	7000	Classification	True Negative	6098
			Average Standard Deviation	6126.9005 18.2127
1	6999	Classification	False Negative	783
2	7000	Classification	False Negative	780
3	7001	Classification	False Negative	791
4	7000	Classification	False Negative	763
5	7000	Classification	False Negative	766
6	7000	Classification	False Negative	771
7	7000	Classification	False Negative	785
8	7000	Classification	False Negative	771
9	7000	Classification	False Negative	772
10	7000	Classification	False Negative	772
			Average Standard Deviation	775.4001 8.476
1	6999	Likelihood	Log Score	-0.3493
2	7000	Likelihood	Log Score	-0.3507
3	7001	Likelihood	Log Score	-0.3488
4	7000	Likelihood	Log Score	-0.3472
5	7000	Likelihood	Log Score	-0.3493
6	7000	Likelihood	Log Score	-0.3517
7	7000	Likelihood	Log Score	-0.349

8	7000	Likelihood	Log Score	-0.3461
9	7000	Likelihood	Log Score	-0.35
10	7000	Likelihood	Log Score	-0.3503
			Average	-0.3492
			Standard Deviation	0.0015
1	6999	Likelihood	Lift	0.0076
2	7000	Likelihood	Lift	0.0062
3	7001	Likelihood	Lift	0.0083
4	7000	Likelihood	Lift	0.0096
5	7000	Likelihood	Lift	0.0076
6	7000	Likelihood	Lift	0.0051
7	7000	Likelihood	Lift	0.0079
8	7000	Likelihood	Lift	0.0107
9	7000	Likelihood	Lift	0.0069
10	7000	Likelihood	Lift	0.0066
			Average	0.0076
			Standard Deviation	0.0016
1	6999	Likelihood	Root Mean Square Error	0.1371
2	7000	Likelihood	Root Mean Square Error	0.1374
3	7001	Likelihood	Root Mean Square Error	0.1277
4	7000	Likelihood	Root Mean Square Error	0.1464
5	7000	Likelihood	Root Mean Square Error	0.1373
6	7000	Likelihood	Root Mean Square Error	0.1368
7	7000	Likelihood	Root Mean Square Error	0.1308
8	7000	Likelihood	Root Mean Square Error	0.135
9	7000	Likelihood	Root Mean Square Error	0.1403
10	7000	Likelihood	Root Mean Square Error	0.1428
			Average	0.1372
			Standard Deviation	0.0051
Decision Tree DMX				

Partition Index	Partition Size	Test	Measure	Value
1	6999	Classification	True Positive	12
2	7000	Classification	True Positive	8
3	7001	Classification	True Positive	5
4	7000	Classification	True Positive	8
5	7000	Classification	True Positive	54
6	7000	Classification	True Positive	7
7	7000	Classification	True Positive	10
8	7000	Classification	True Positive	9
9	7000	Classification	True Positive	49
10	7000	Classification	True Positive	43
			Average Standard Deviation	20.4999 18.6829
1	6999	Classification	False Positive	1
2	7000	Classification	False Positive	1
3	7001	Classification	False Positive	0.000e+000
4	7000	Classification	False Positive	1
5	7000	Classification	False Positive	117
6	7000	Classification	False Positive	1
7	7000	Classification	False Positive	1
8	7000	Classification	False Positive	1
9	7000	Classification	False Positive	117
10	7000	Classification	False Positive	116
			Average Standard Deviation	35.6 53.072
1	6999	Classification	True Negative	6193
2	7000	Classification	True Negative	6194
3	7001	Classification	True Negative	6195
4	7000	Classification	True Negative	6194
5	7000	Classification	True Negative	6078
6	7000	Classification	True Negative	6194
7	7000	Classification	True Negative	6194
8	7000	Classification	True Negative	6194
9	7000	Classification	True Negative	6078
10	7000	Classification	True Negative	6079
			Average Standard Deviation	6159.3 53.0077
1	6999	Classification	False Negative	793

2	7000	Classification	False Negative	797
3	7001	Classification	False Negative	801
4	7000	Classification	False Negative	797
5	7000	Classification	False Negative	751
6	7000	Classification	False Negative	798
7	7000	Classification	False Negative	795
8	7000	Classification	False Negative	796
9	7000	Classification	False Negative	756
10	7000	Classification	False Negative	762
			Average Standard Deviation	784.6001 18.7681
1	6999	Likelihood	Log Score	-0.3426
2	7000	Likelihood	Log Score	-0.3425
3	7001	Likelihood	Log Score	-0.3458
4	7000	Likelihood	Log Score	-0.3423
5	7000	Likelihood	Log Score	-0.3451
6	7000	Likelihood	Log Score	-0.3442
7	7000	Likelihood	Log Score	-0.3411
8	7000	Likelihood	Log Score	-0.3408
9	7000	Likelihood	Log Score	-0.3472
10	7000	Likelihood	Log Score	-0.3438
			Average Standard Deviation	-0.3435 0.002
1	6999	Likelihood	Lift	0.0143
2	7000	Likelihood	Lift	0.0144
3	7001	Likelihood	Lift	0.0113
4	7000	Likelihood	Lift	0.0145
5	7000	Likelihood	Lift	0.0117
6	7000	Likelihood	Lift	0.0127
7	7000	Likelihood	Lift	0.0158
8	7000	Likelihood	Lift	0.016
9	7000	Likelihood	Lift	0.0096
10	7000	Likelihood	Lift	0.013
			Average Standard Deviation	0.0133 0.0019
1	6999	Likelihood	Root Mean Square Error	0.1245
2	7000	Likelihood	Root Mean Square Error	0.1255

3	7001	Likelihood	Root Mean Square Error	0.1247
4	7000	Likelihood	Root Mean Square Error	0.1244
5	7000	Likelihood	Root Mean Square Error	0.1589
6	7000	Likelihood	Root Mean Square Error	0.1241
7	7000	Likelihood	Root Mean Square Error	0.1247
8	7000	Likelihood	Root Mean Square Error	0.124
9	7000	Likelihood	Root Mean Square Error	0.1563
10	7000	Likelihood	Root Mean Square Error	0.1555
			Average	0.1343
			Standard Deviation	0.0149

Comparing Root Mean Square error, Lift and log score for all the three models,

S.No	Model Name	RMSE	Lift Score	Log Score
1	Decision Tree	0.1343	0.0133	-0.3435
2	Logistic Regression	0.1449	0.0089	-0.348
3	Neural Networks	0.1372	0.0076	-0.3492

We see from the comparison that Decision Tree performed better in terms of RMSE, Lift and Log score.

Summary of Results:

Among the three models Decision tree performed better in terms of predicting the serious injury in accidents.

Hierarchy of variable importance in predicting the accident severity is as follows,

1) Vehicle Type
2) Junction Details
3) Urban rural
4) Speed Limit

If a person, who is a male and is traveling in car with a speed greater than 60 is prone to severe injury.

The Decision Tree will correctly identify 46.21% of the people who are seriously injured in the entire population. To identify the severe injury among the people , we would use query to retrieve cases with a predict probability of 12.96%

Conclusion:

The decision tree made sensible prediction in determining the serious injury factors. People shouldn’t over speed in a car to avoid accident.