UK ACCIDENTS   
 DATABASE

Data warehousing project under the guidance of Dr. Rathin Sarathy

Table of Contents:

Cube Creation ……………………………………………………………………………………3

Named Calculations Creation …………………………………………………………….6

Measures ………………………………………………………………………………………….10

Hierarchy ………………………………………………………………………………………….12

Cube Partitions and aggregations …………………………………………………….19

MDX Queries ……………………………………………………………………………………19

Data Mining …………………………………………………………………………………….26

Creating Mining Structure ……………………………………………………………….36

Creating Mining Models ………………………………………………………………….37

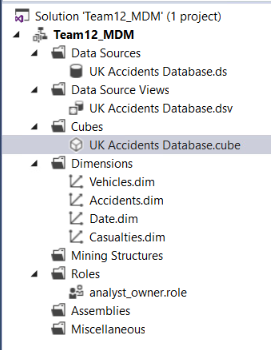
Comparison of Results …………………………………………………………………….48

Summary …………………………………………………………………………………………55

Conclusion ………………………………………………………………………………………56

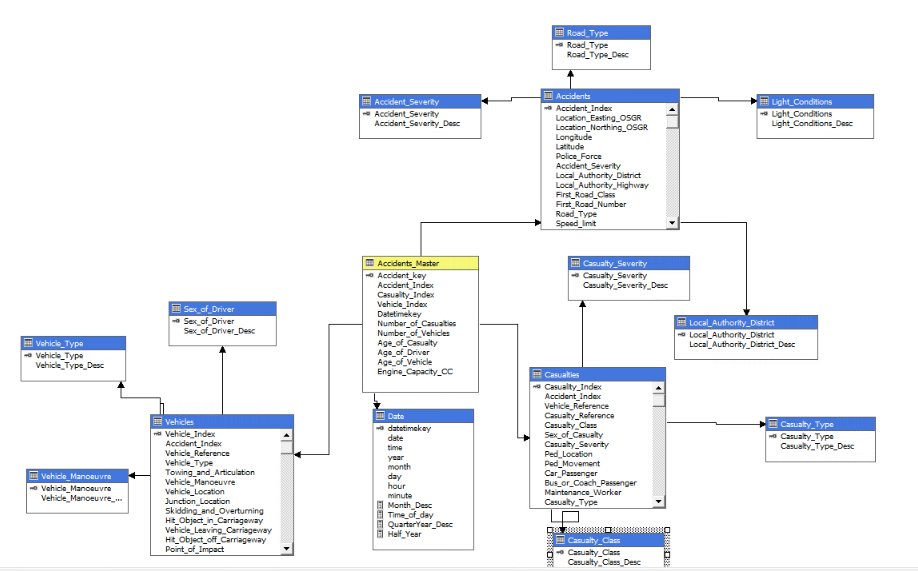
**Cube Creation**

Step1: We must first create a connection to the data source in visual studio. i.e. we will connect to the relational database called UK Accidents Database.   
  
The Team12\_MDM is our new analysis database.   
UK Accidents Relational Database will act as the source of data for our Team12\_MDM analysis database.

****

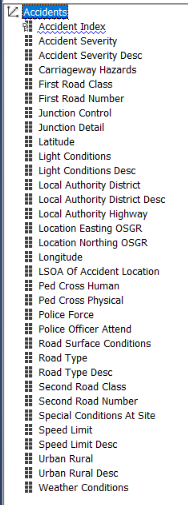
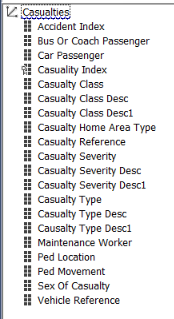
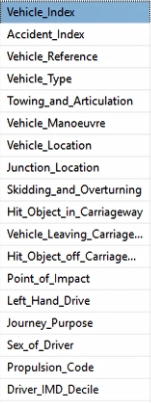
We see the new data source being connected to the UK Accidents relational database. Later, we deployed the data source to SQL Server Analysis Services.

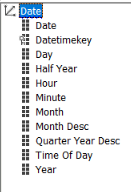
Step 2: Create the data source view  
  
 We created the data source view which is the link between UK accidents relational database and the analysis services database Team12\_MDM.   
  
While creating the New Data Source View we had the option to include Accidents Master fact table. Later, we added the related tables. After the DSV is created within visual studio, it must be deployed to view the DSV within SQL Server Management Studio. This DSV becomes / acts as the relational database.

****

The above screenshot shows the DSV created within visual studio. It shows all the Dimension tables and Accident\_Master Fact table.  
  
We used the DimAccidents (renamed as Accidents) table within the DSV to create the analysis services Accidents Dimension.  
  
Similarly, we used DimVehicles, DimDate, DimCasulaties tables within DSV (all renamed as shown in the above screenshot) and created the analysis services Vehicle, Date and Casualties Dimensions.

Accidents Dimension Casualties Dimension Vehicle Dimension

**  **  
Date Dimension



**Named Calculations Creation**

Named calculations refer to the new attributes created within the table. We use T-SQL since the DSV tables are relational databases.

* **Month\_desc: is the concatenation of month and year**

CASE

WHEN month = 1 THEN CONCAT('January',' ' ,year)

WHEN month = 2 THEN CONCAT('February',' ' ,year)

WHEN month = 3 THEN CONCAT('March',' ' ,year)

WHEN month = 4 THEN CONCAT('April',' ' ,year)

WHEN month = 5 THEN CONCAT('May',' ' ,year)

WHEN month = 6 THEN CONCAT('June',' ' ,year)

WHEN month = 7 THEN CONCAT('July',' ' ,year)

WHEN month = 8 THEN CONCAT('August',' ' ,year)

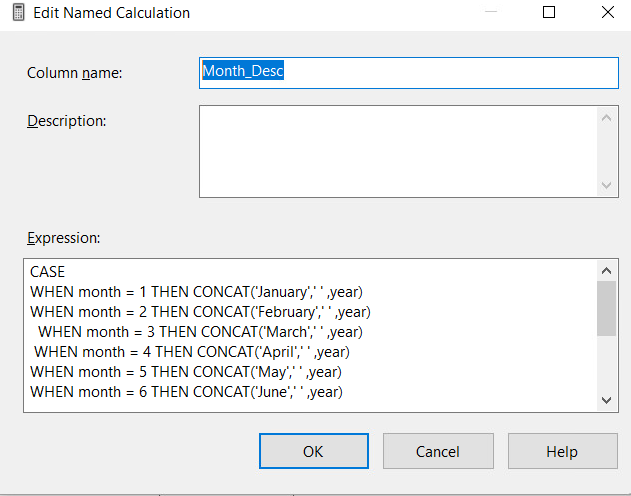
WHEN month = 9 THEN CONCAT('September',' ' ,year)

WHEN month = 10 THEN CONCAT('October',' ' ,year)

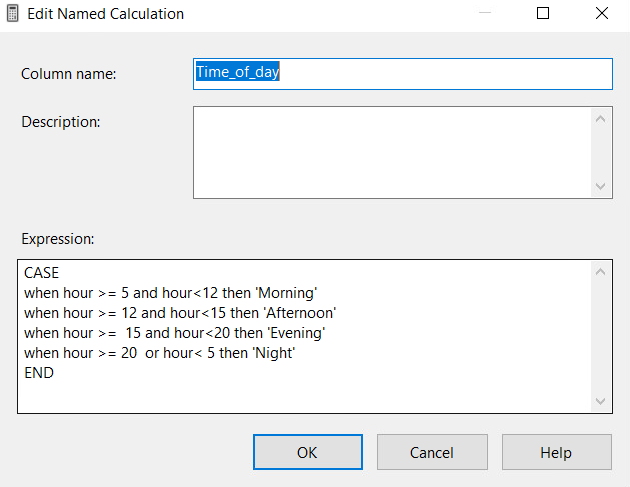
WHEN month = 11 THEN CONCAT('November',' ' ,year)

WHEN month = 12 THEN CONCAT('December',' ' ,year)

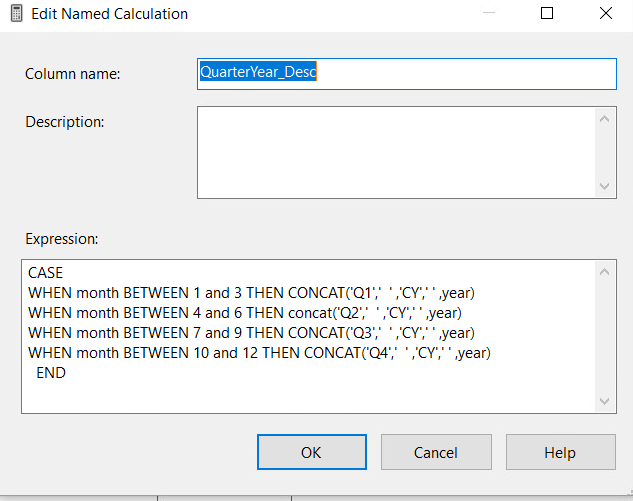
END



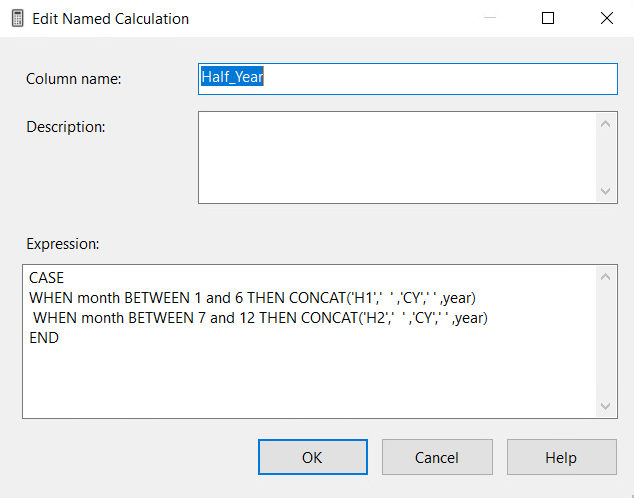
* **Time\_of\_day** : which explains different parts of the day



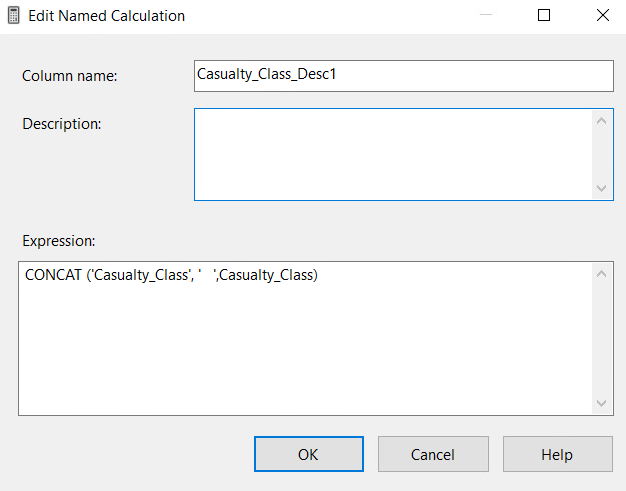
* **QuaterYear\_Desc** : which categorizes months into each quarter



* **Half\_Year**: which categories quarters into semesters

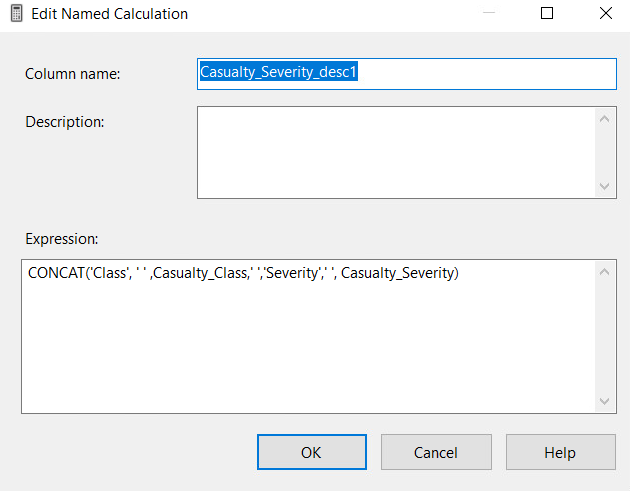


* **Casualty\_Class\_Desc1**: comprises of casualty class



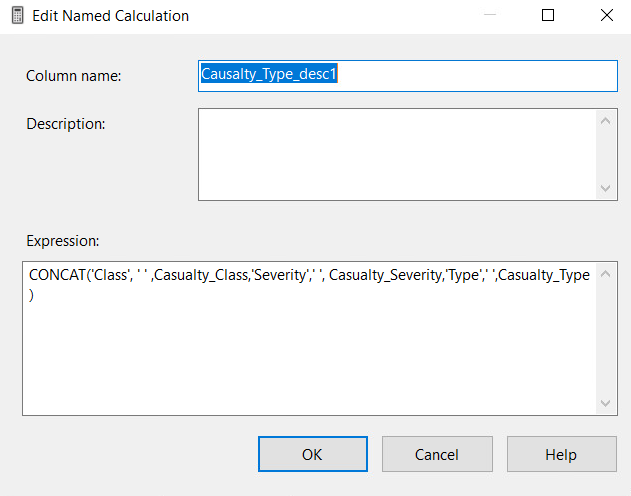
* **Casualty\_Severity\_desc1**:is the concatenation of casualty\_class and casualty\_severity

i.e. Every casualty class has 3 casualty severities. In the below diagram, we are   
 concatenating casualty class with casualty severity to get a unique record.



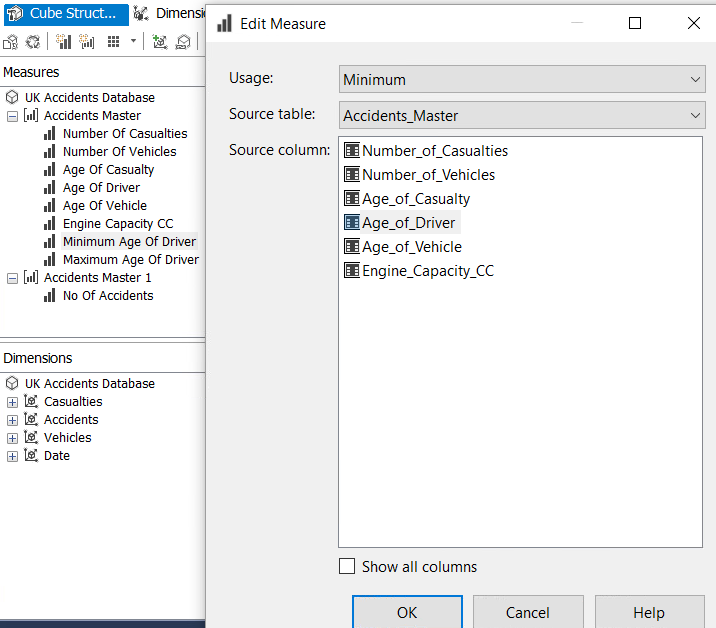
* **Casualty\_Type\_desc1** : is the concatenation of casualty\_class, casualty\_severity, and casualty\_type

i.e. Every casualty class has 3 casualty severities. Every casualty severity has further   
 types. In the below diagram, we are concatenating casualty class with casualty severity   
 and casualty type to get a unique record.

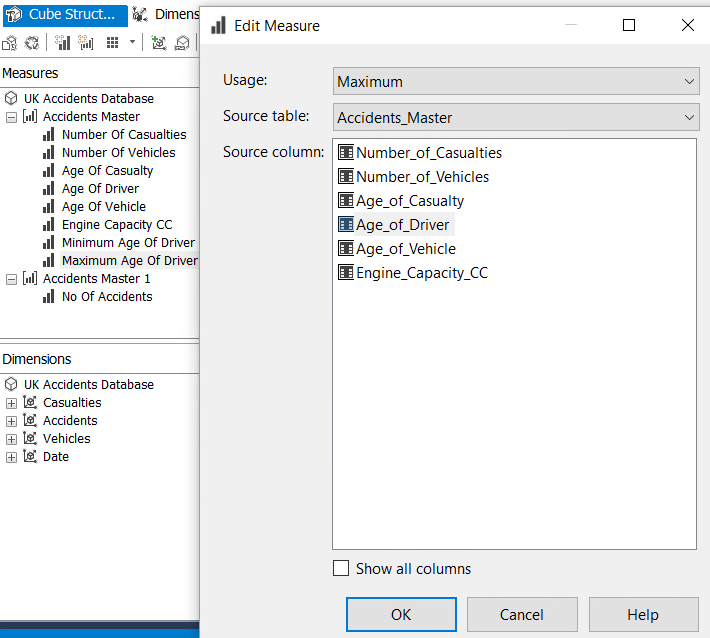


**New Measures**

* **Minimum Age Of Driver** Measure

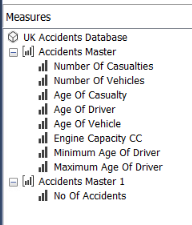


* **Maximum Age Of Driver** Measure



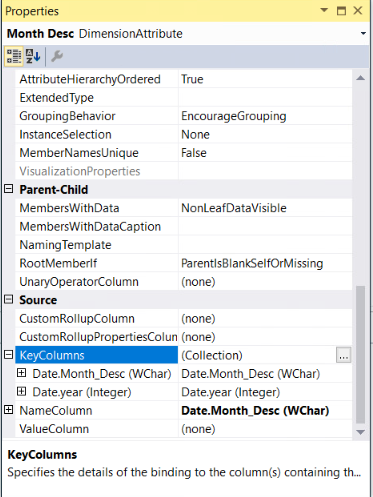
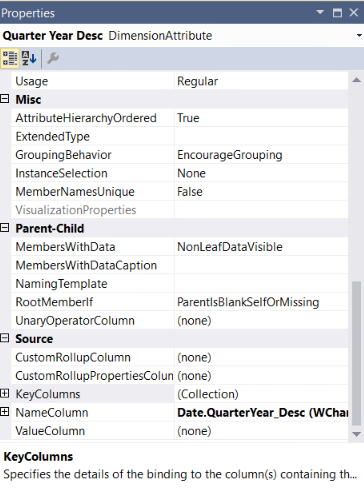
* **No of accidents Measure**

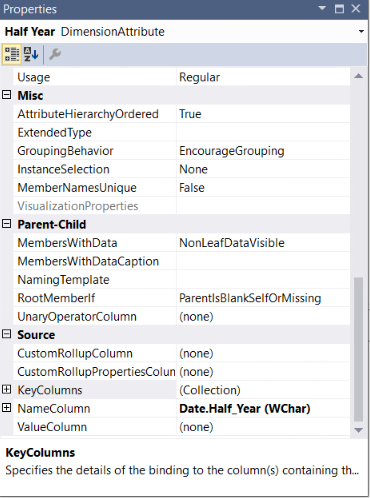
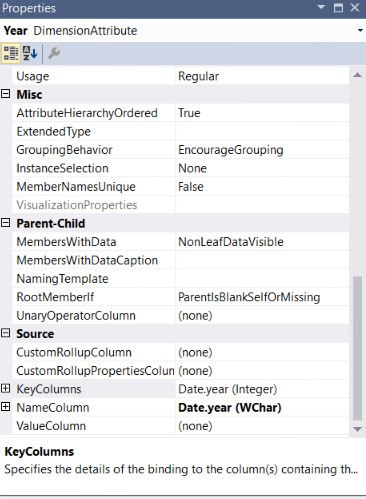


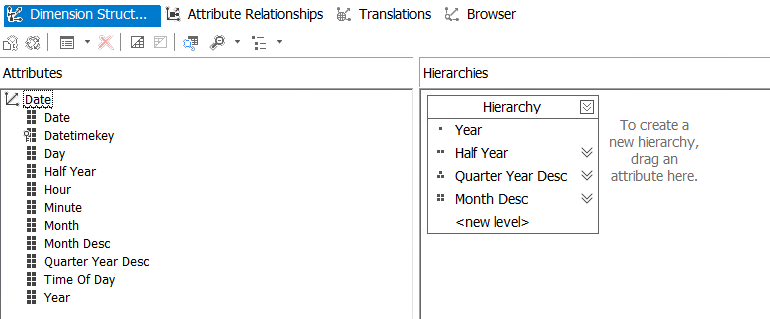
****

The above screenshot shows all the 3 measures created.

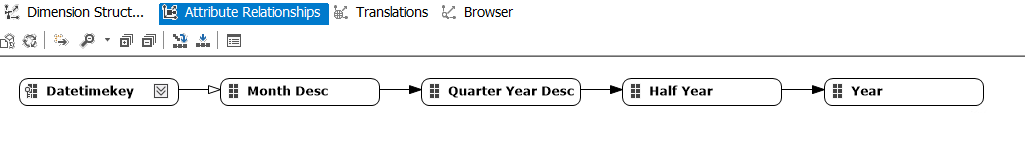
**Hierarchy**  
   
Hierarchies permit us to drill across, drill-down and roll up aggregates.  
We need to fix the keycolum and name column values found within the   
attributes of the dimension.  
  
**Date Hierarchy Key columns and named columns fixing**  
In-order to have unique values for a month description , we used a composite key which   
contains month within a given year. For instance, January month is available across all   
the years in the data source we have. To have a unique record of January, we concatenate   
Jan with CY and Year. i.e. the month name is not January but January CY 2015 . Similarly   
we have set composite keycolumns for Quarters and half-year as shown below

** **

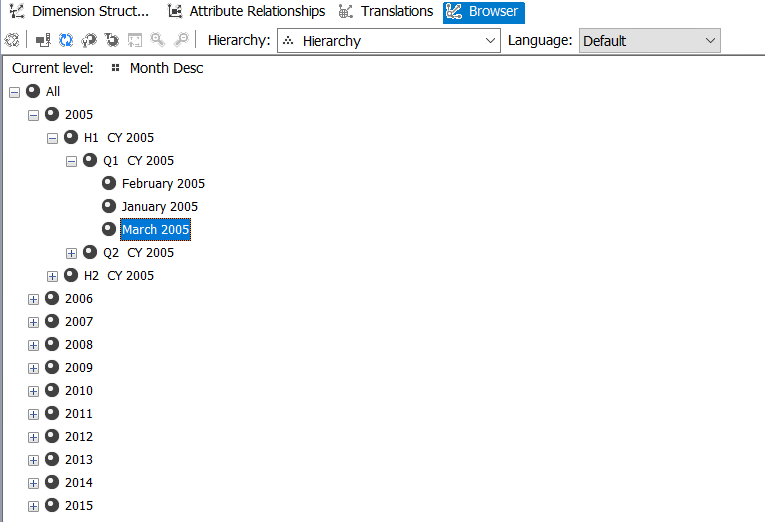
** **



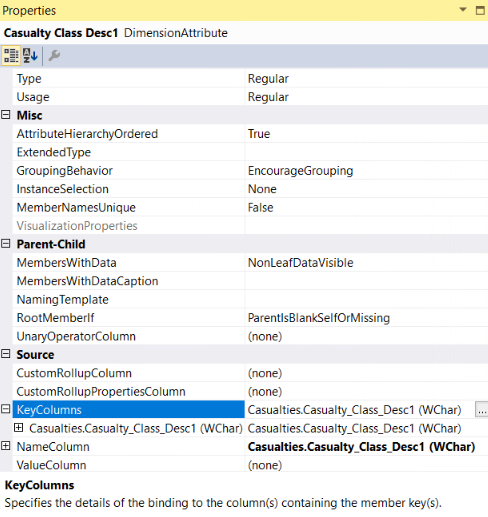
The attribute relationships for date dimension hierarchy are shown below

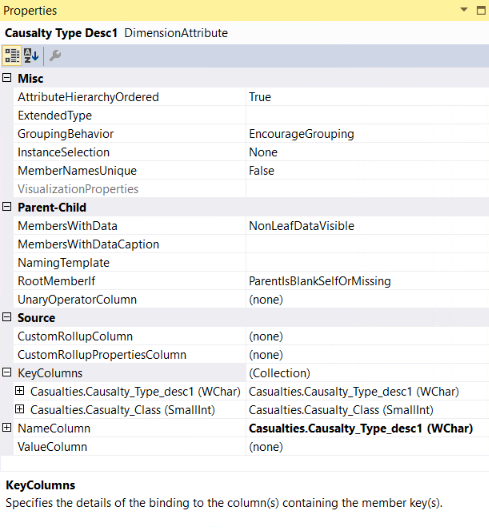
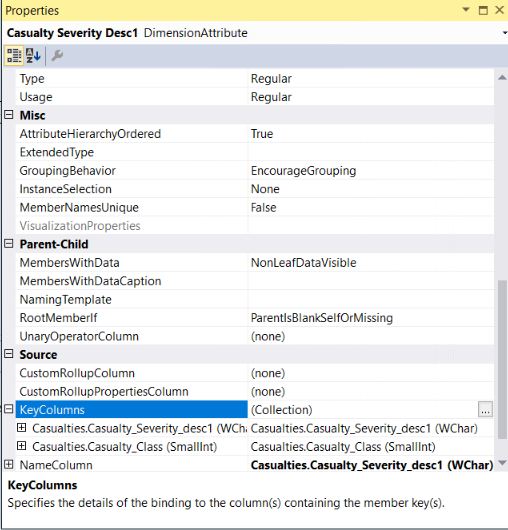


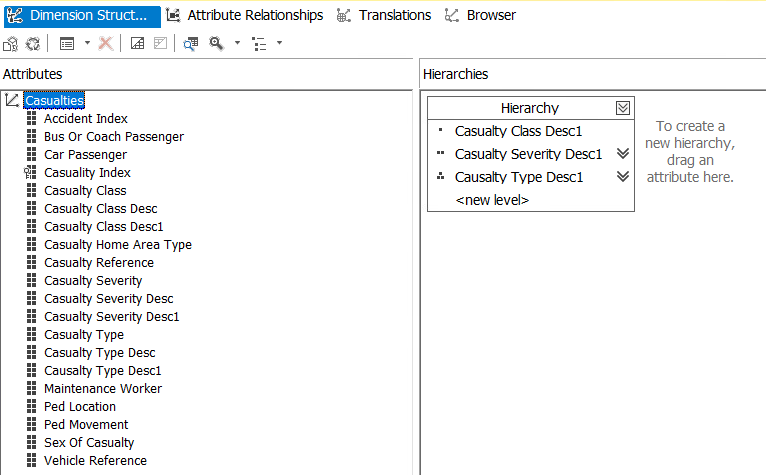
In the below date hierarchy screenshot, we see the individual years from 2005 to 2015. We can drill down within any given year to find the half yearly, quarterly and monthly records respectively.

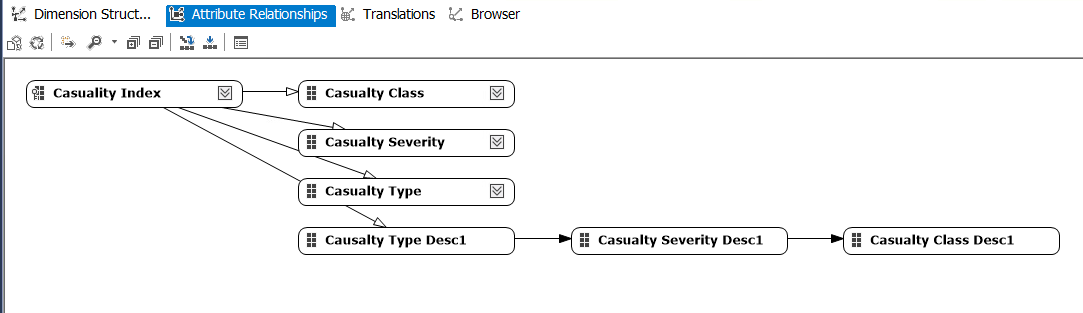


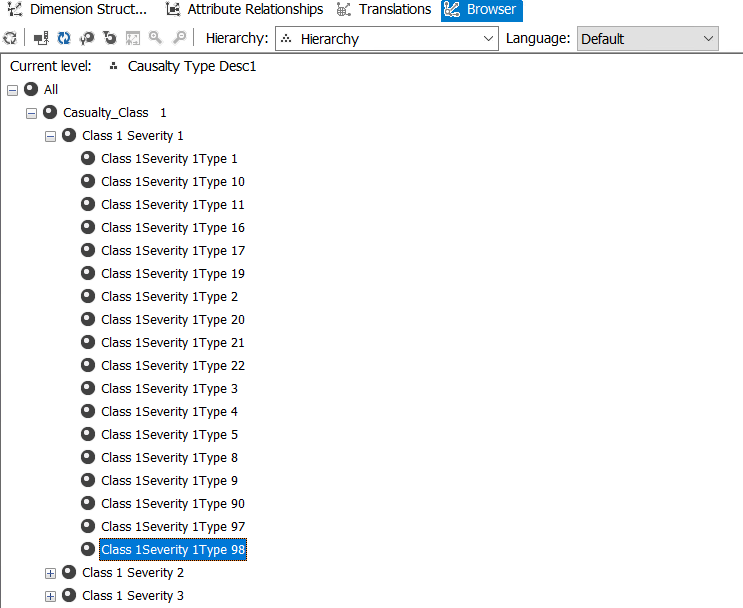
**Casualties Dimension Hierarchy**Every casualty class has 3 casualty severities. Every casualty severity has further casualty   
 types. To get a unique record for each casualty class, severity and type we have a composite key structure as shown below.

****

** **

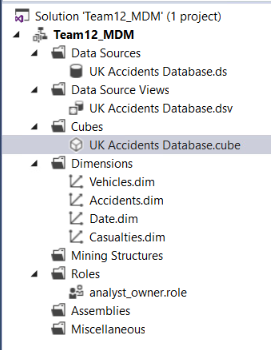




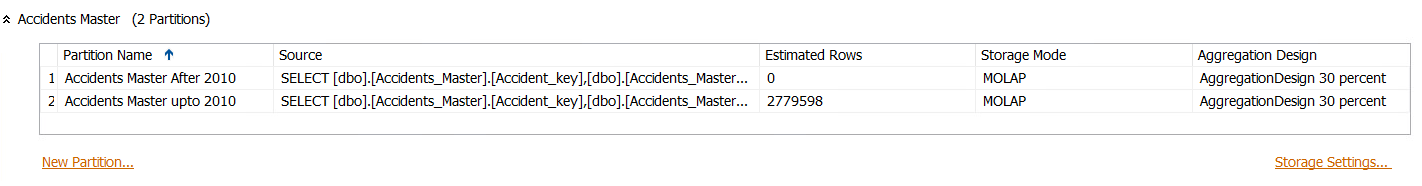


**UK Accidents Database Cube**

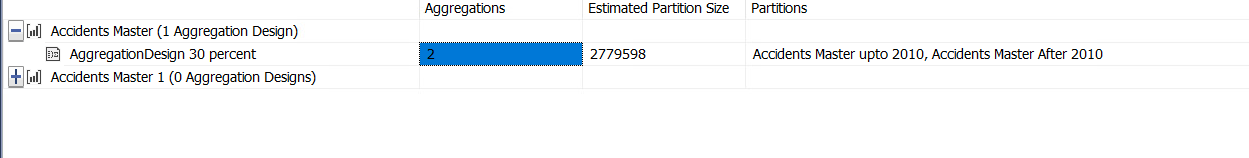
We built the UK Accidents Database Cube using Vehicles, Accidents,date and Casualties dimensions.

****

**Partitions and Aggregation**  
  
Partitions are used to manage and store data and aggregations for a measure group within a cube. When we process a partition, data is brought into the partition from the source. ( Reference: Lecture 6 Part B slide 25)  
We have created and deployed 2 partitions of the Accidents Masters Fact table( datetimekey – one up to year 2010 , the other after 2010).   
The 1st partition includes any dates less than 2010 and the 2nd partition is based on the dates greater than 2010.

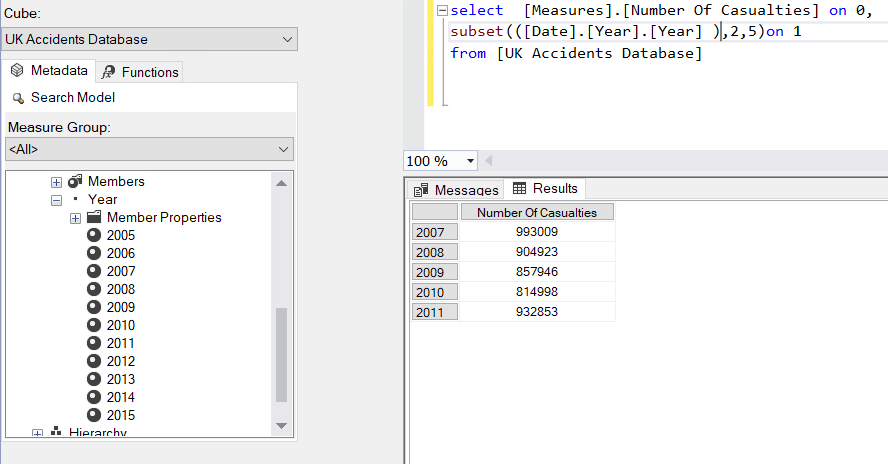


We created Aggregations in each partition for 30% Performance.

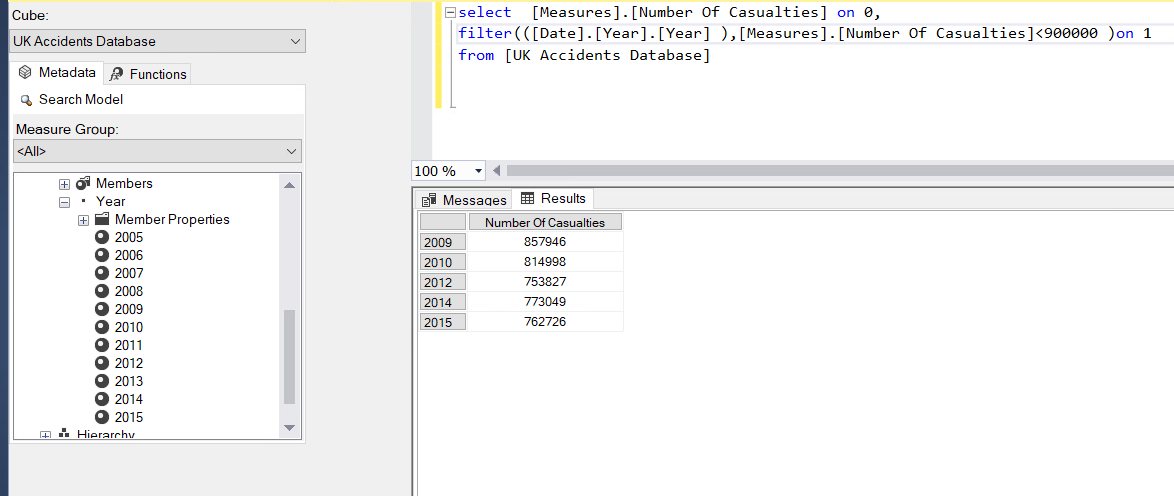


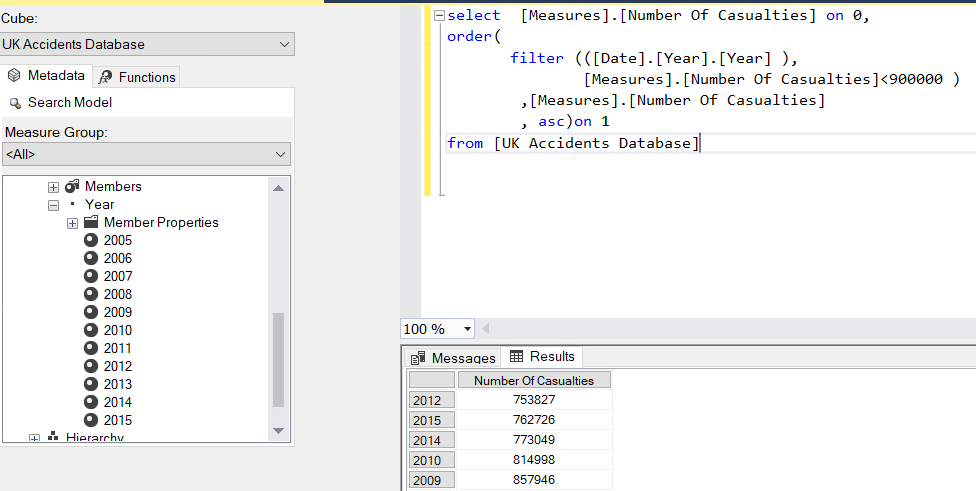
**MDX Queries**

**1 Display the number of casualties for the next 5 years starting from second year in the   
 date hierarchy**

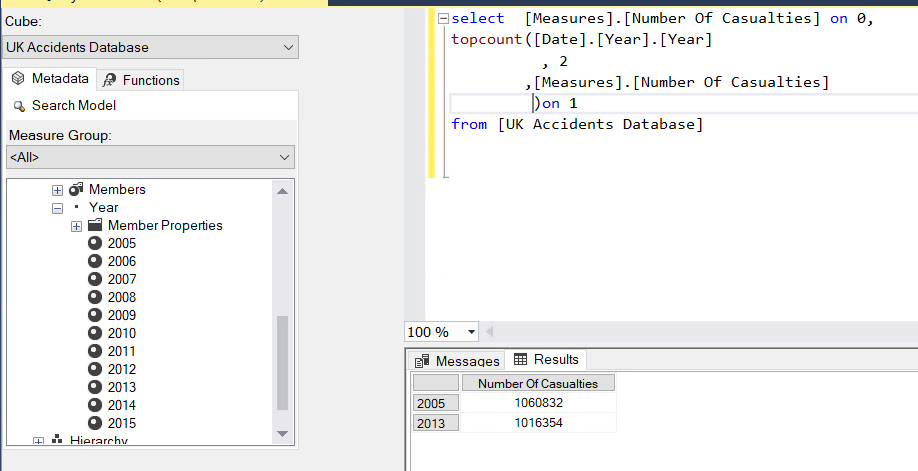


**2. Display the number of Casualties for the years where number of casualties are less than 900000**

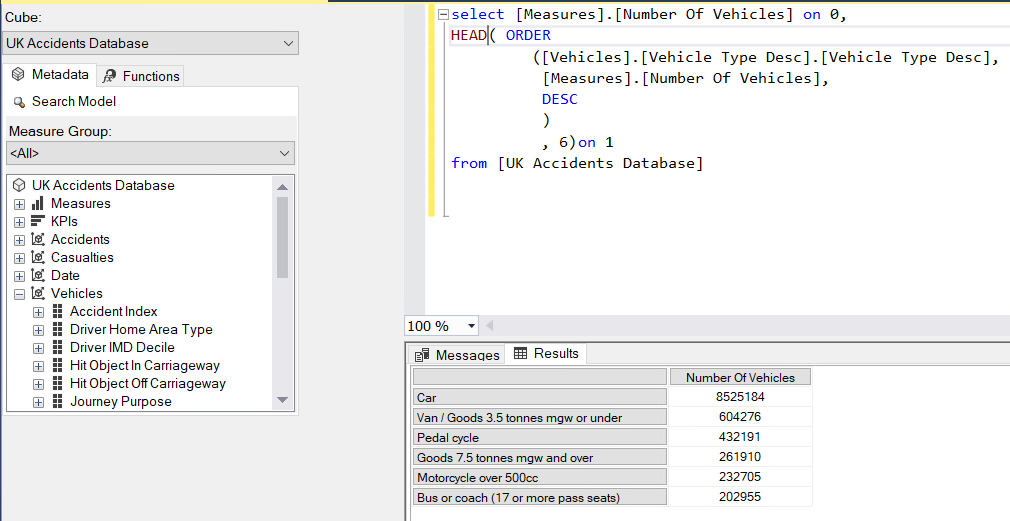


**3. After displaying the casualties in the previous question, order by the number of casualties (ascending order)**

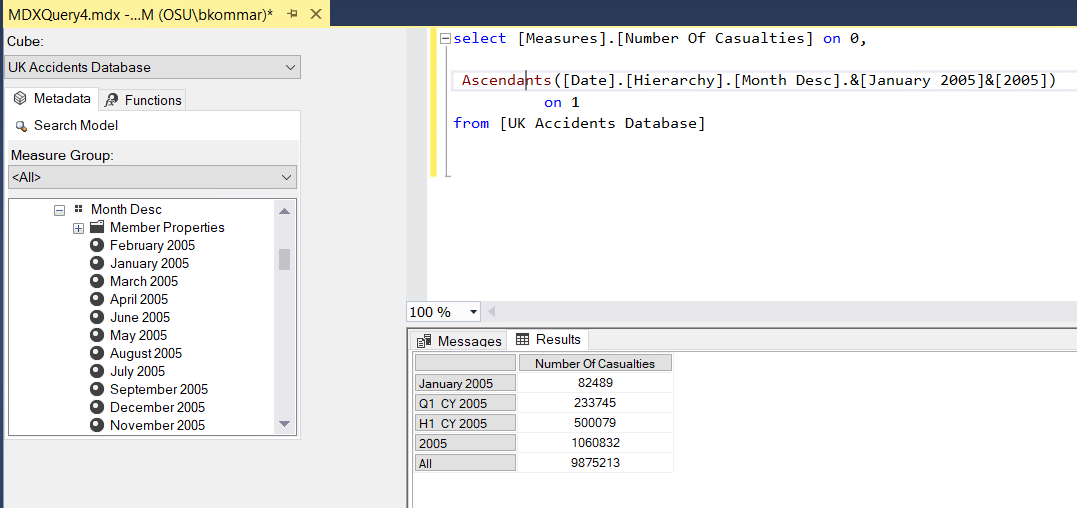
**4. Display the top 2 years with the highest number of casualties**



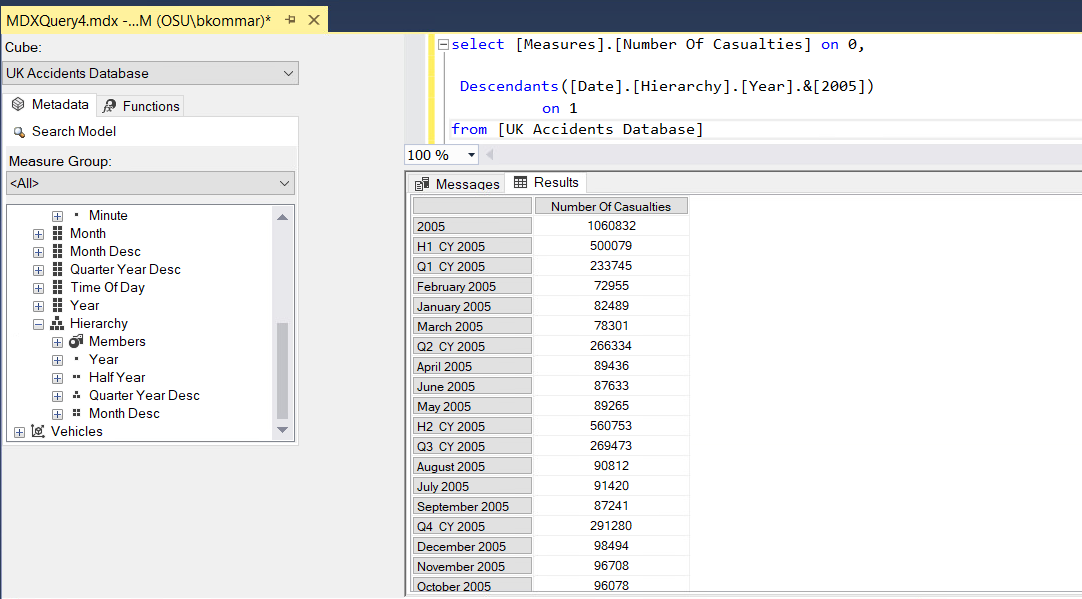
**5. Display the top 6 vehicles and the vehicle description**



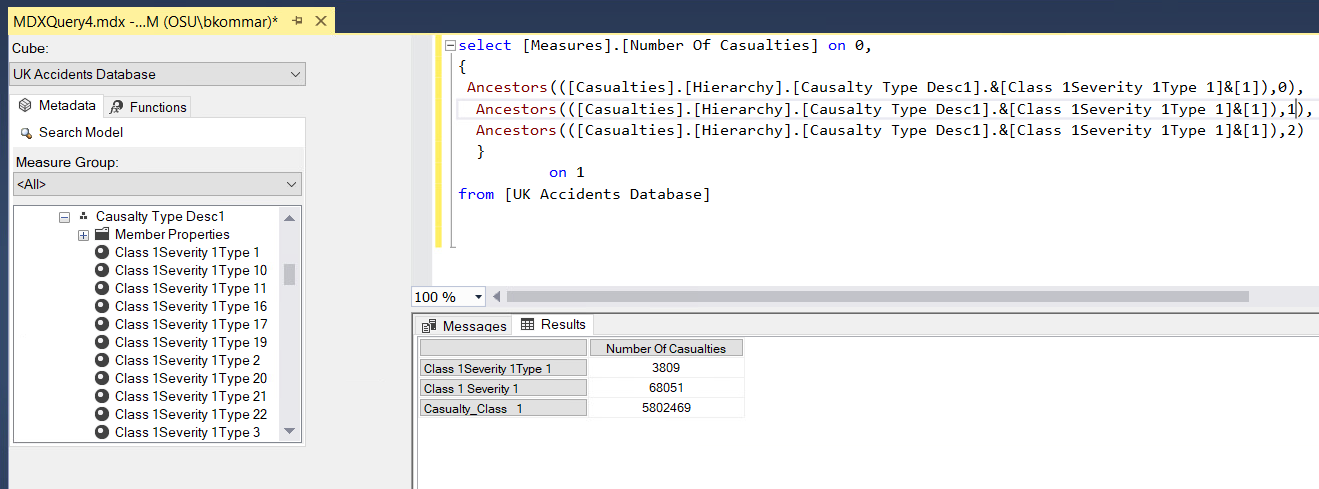
**6. Display the number of casualties for the ascendants in the year 2005**

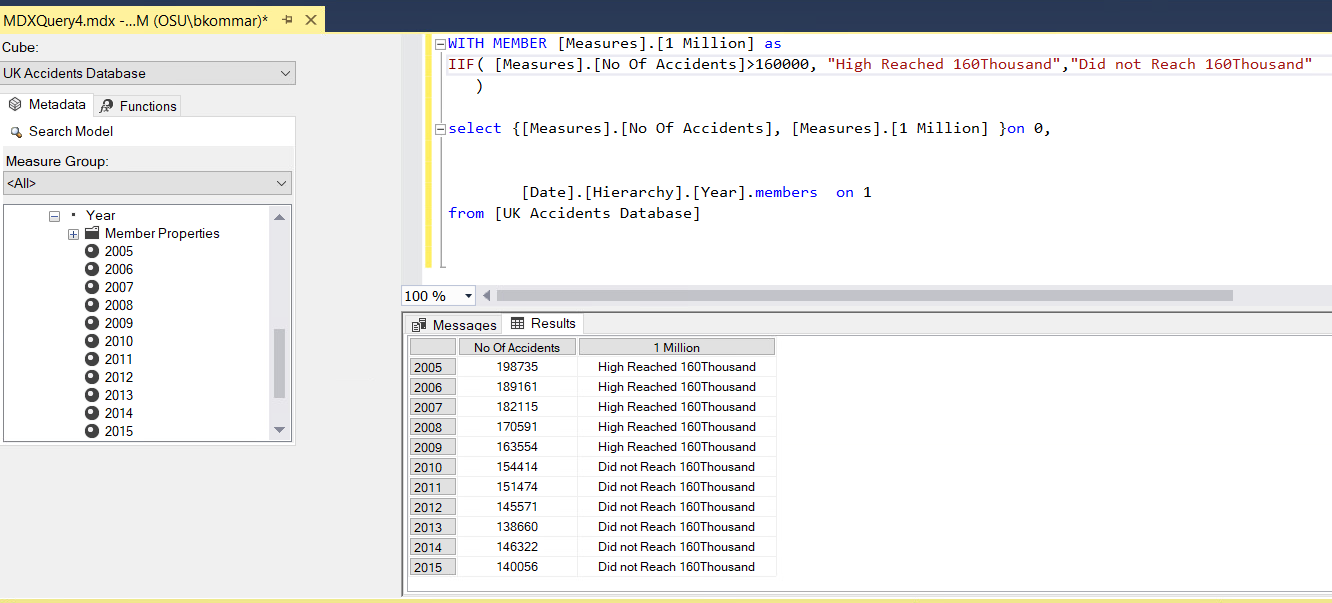


**7. Display the number of casualties for the descendants in the year 2005**



**8. Display the number of casualties for the ancestors of Class1Severity1Type1 in the casualty hierarchy (user-defined) at level 0,1,2**

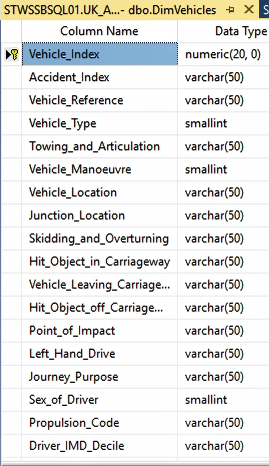
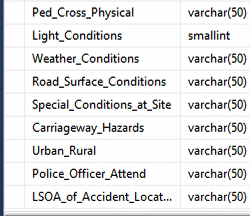


**9. Display the number of accidents if it reached 160,000 or not for every year** 

**10. Display the number of casualties for all years using GENERATE function.** A screenshot of a social media post

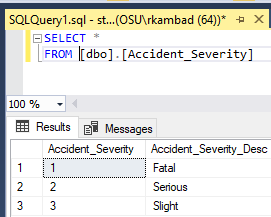
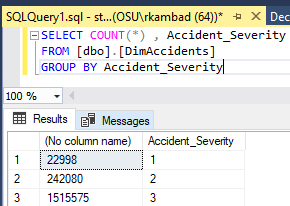
Description generated with very high confidence

**Data Mining**   
  
Data mining involves exploring and analyzing large data to discover the hidden patterns and rules. It’s basically a technique used to predict future outcomes. ([DataMining](https://www.microstrategy.com/us/resources/introductory-guides/data-mining-explained))

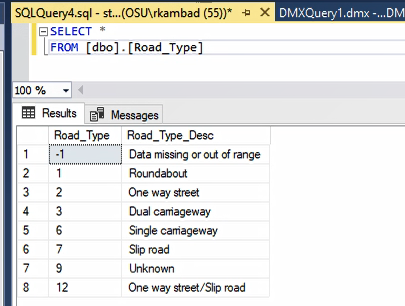
    
 

The above pictures show the attributes of Accidents and Vehicles tables. We decided to predict accident severity. To decide on the contributing factors for accident severity, it was necessary to learn more about each attribute in both the above tables.

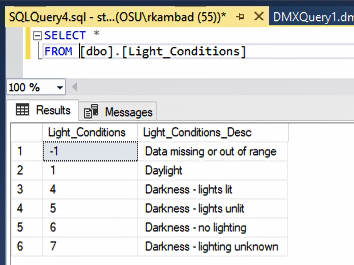
**Accident Severity** (we will be predicting this)  
  
This is the target variable or the variable we would be predicting. As seen below, the accident severity which is ‘Slight’ has highest number of records within accidents table.

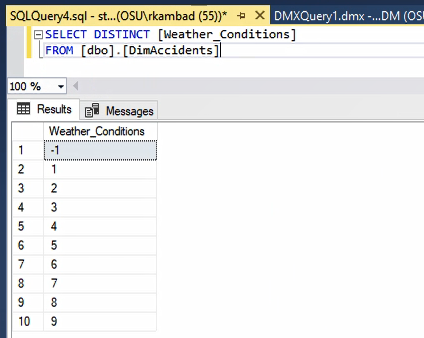
**Road\_Type**  
  
We understand that road type plays an important role in predicting the severity of an accident. To find out what each road type numerical value within the Accidents table meant, we queried the Road\_Type table.   
  
As seen below, we will consider all the road type values except road type = -1 or road type = 9 to predict the accident severity.



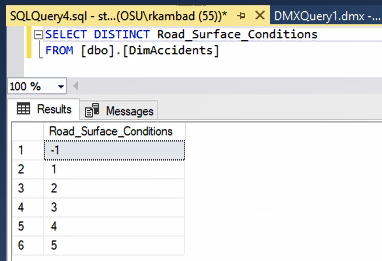
**Light condition**  
Light condition is another factor which greatly influences driving. Darkness while driving leads to accidents. To find out what each numerical value of light\_conditions within the Accidents table meant, we queried the Light\_conditions table.   
  
As seen below, we will consider all the road type values except road type = -1 to predict the accident severity

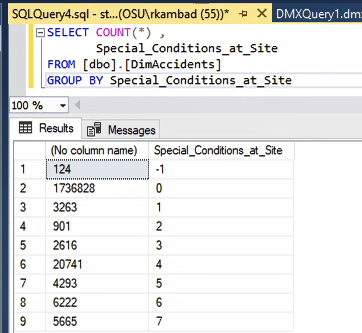


**Weather Conditions**  
  
Weather also helps in predicting accidents. Although we didn’t have any table that specifically explains what each numerical value of weather conditions within the Accidents table meant, we decided to consider all numerical values except weather\_condition =- 1 (believing it to be unknown or missing values) to predict accident severity.

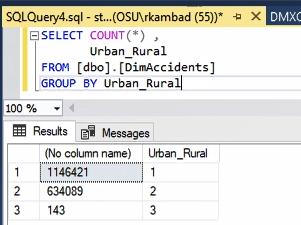


**Road Surface Condition**  
  
The condition of roads may also help in predicting accident severity. We decided to consider all the values except road\_surface\_conditions = -1 for predicting accident severity.

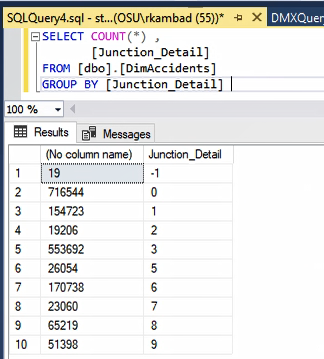


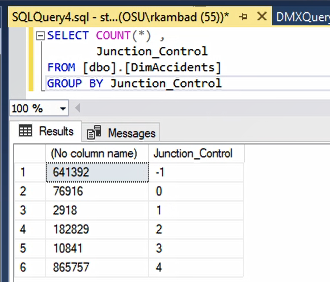
**Special Conditions at Site**  
  
We decided to eliminate special conditions at site = -1 while predicting accident severity. Although we do not know what special\_conditions\_at\_site = 0 meant, based on the large number of records available, we decided to consider this value also for predicting accident severity.  


**Urban Rural**   
  
We know that accidents happen more in urban areas compared to rural as the number of vehicles operating in urban areas are more. We think that the value of 1 indicates urban and a 2 indicates rural. Although there are only 143 records for urban\_rural =3 within the accidents table, we decided to use all the three values of urban\_rural to predict accident severity.

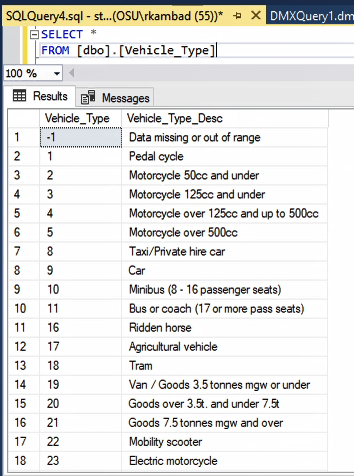
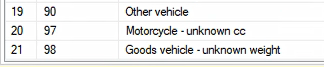


**Junction detail**   
  
A junction is where two or more roads meet. It may also influence the accident severity. Hence, we are using this variable within our mining structure. We will not consider junction\_detail = -1 for predicting accident severity.



**Junction Control**   
  
We will not consider junction\_control=-1 for predicting the accident severity.

**Vehicle Type**  
  
The type of vehicle whether car, cycle and a motorcycle etc may also help in predicting accident severity.

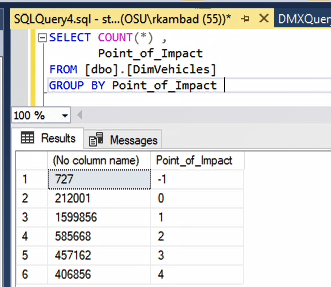
 

**Vehicle Manoeuvre**  
  
To understand what each numerical value of vehicle\_manoeuvre meant within the accidents table,we queried the Vehicle Manoeuvre table. The description of the values can be seen in the below screenshot. Whether a vehicle was taking a U-turn, overtaking another or changing lanes may also help in predicting accident severity.  
  
We won’t consider vehicle\_manoeuvre = -1 for our mining models.

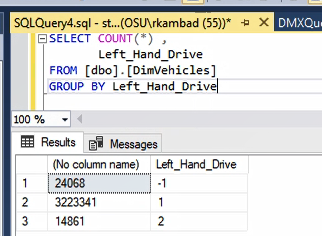


**Point of Impact**

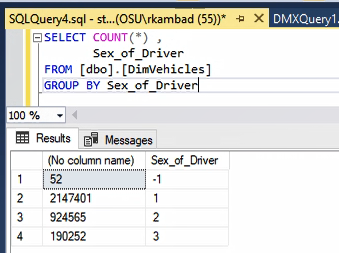
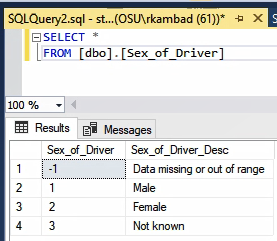
We think that the point of impact refers to whether the impact was on the left, right, front or rear of the vehicle etc. We will consider all the values for point of impact except -1



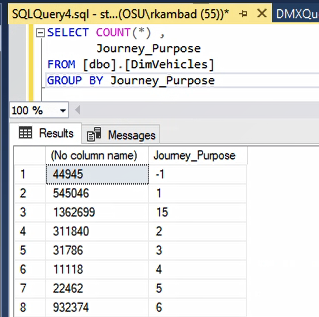
**Left Hand Drive**  
  
We will not consider the left\_hand\_drive = -1 for our mining models.



**Sex of Driver**  
  
The sex of the driver may also help in predicting accident severity. We have considered only the values sex of driver = 1 or 2 to predict the severity of accidents.

**Journey Purpose**  
  
Journey purpose may also help in predicting the severity of an accident. We will not consider journey\_purpose = -1 for our mining model.



Based on the above analysis of the attributes within Accidents and Vehicle table, we used the following attributes to predict accident severity.

Attributes from accidents table

* Accident\_Index - Key column which uniquely identifies an entity
* Road\_Type
* Speed\_limit
* Light\_Conditions
* Weather\_Conditions
* Road\_Surface\_Conditions
* Special\_Conditions\_at\_Site
* Urban\_Rural
* Junction\_Detail
* Junction\_Control

Attributes from Vehicles table

* Vehicle\_Type
* Vehicle\_Manoeuvre
* Point\_of\_Impact
* Left\_Hand\_Drive
* Sex\_of\_Driver
* Journey Purpose

Various mining models could be built using the above UK Accidents Database. We decided to use the following data mining techniques based on Microsoft data mining algorithms.

* Decision Tree
* Logistic Regression
* Neural Networks

**Creating the Mining Structure**   
  
We will use the CREATE MINING STRUCTURE DMX statement to create the mining structure. Since we will be creating many mining models, we will use ALTER MINING STRUCTURE statement to add mining models to the structure.   
  
The name of the mining structure we created is : Accident Severity DMX

CREATE MINING STRUCTURE [Accident Severity DMX]

( [Accident\_Index] LONG KEY

,[Accident\_Severity] LONG DISCRETE /\*predicting\*/

,[Road\_Type] LONG DISCRETE

,[Speed\_limit] TEXT DISCRETE

,[Light\_Conditions] LONG DISCRETE

,[Weather\_Conditions] TEXT DISCRETE

,[Road\_Surface\_Conditions] TEXT DISCRETE

,[Special\_Conditions\_at\_Site] TEXT DISCRETE

,[Urban\_Rural] TEXT DISCRETE

,[Junction\_Detail] TEXT DISCRETE

,[Junction\_Control] TEXT DISCRETE

,[Vehicle\_Type] LONG DISCRETE

,[Vehicle\_Manoeuvre] LONG DISCRETE

,[Point\_of\_Impact] TEXT DISCRETE

,[Left\_Hand\_Drive] TEXT DISCRETE

,[Sex\_of\_Driver] LONG DISCRETE

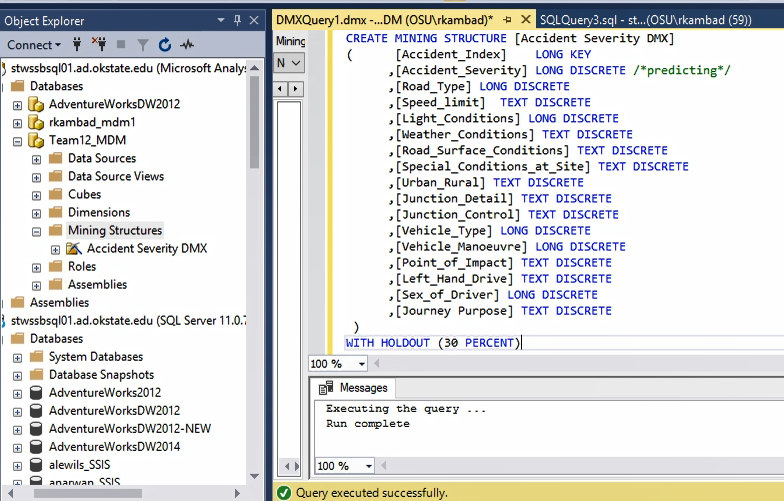
,[Journey Purpose] TEXT DISCRETE

)

WITH HOLDOUT (30 PERCENT)

The key column for the mining structure uniquely identifies an entity in the source data.  
We have also defined the mining columns. Additionally, we specified what portion of the data is used for testing mining models. The remaining data is used for training the models.  
  
By default, analysis services will create a test data set which contains 30% of all the data. We can also add a specification that the test data set should contain 30% of the cases up to a maximum of 1000 cases. (reference lecture 7, slide no. 17)

In the below screenshot, we see the Accident Severity DMX mining structure created under the Mining Structure folder.



**Creating Mining Models**

After identifying the mining structure, we will use the ALTER MINING STRUCTURE STATEMENT to alter the mining structure and add mining models.

It is necessary to define the predictable and the input columns. Additionally, we must determine which algorithm to use.

As stated earlier, we will use the following data mining techniques based on Microsoft data mining algorithms.

* Decision Tree
* Logistic Regression
* Neural Networks

**Model1: Decision Tree Mining Model**

ALTER MINING STRUCTURE[Accident Severity DMX]

ADD MINING MODEL [Decision Tree DMX]

( [Accident\_Index]

,[Accident\_Severity] PREDICT

,[Road\_Type]

,[Speed\_limit]

,[Light\_Conditions]

,[Weather\_Conditions]

,[Road\_Surface\_Conditions]

,[Special\_Conditions\_at\_Site]

,[Urban\_Rural]

,[Junction\_Detail]

,[Junction\_Control]

,[Vehicle\_Type]

,[Vehicle\_Manoeuvre]

,[Point\_of\_Impact]

,[Left\_Hand\_Drive]

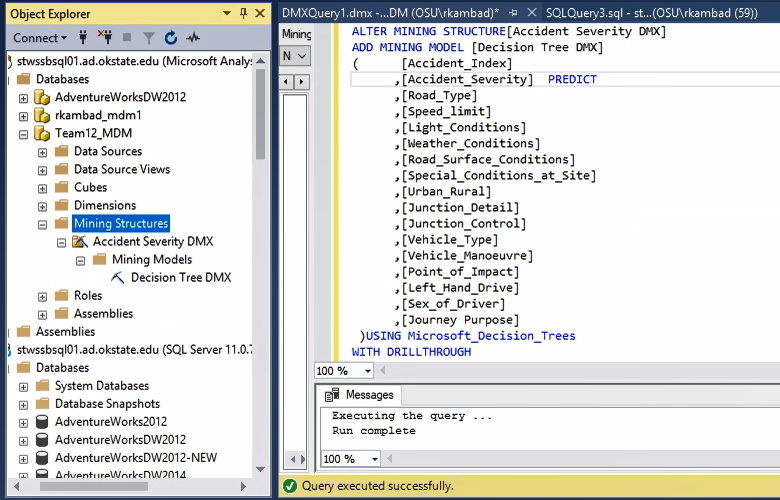
,[Sex\_of\_Driver]

,[Journey Purpose]

)USING Microsoft\_Decision\_Trees

WITH DRILLTHROUGH

In the below screenshot, we see the Decision Tree DMX mining model created under Accident Severity DMX mining structure.



**Model2: Logistic Regression Mining Model**

ALTER MINING STRUCTURE[Accident Severity DMX]

ADD MINING MODEL [Logistic Regression DMX]

( [Accident\_Index]

,[Accident\_Severity] PREDICT

,[Road\_Type]

,[Speed\_limit]

,[Light\_Conditions]

,[Weather\_Conditions]

,[Road\_Surface\_Conditions]

,[Special\_Conditions\_at\_Site]

,[Urban\_Rural]

,[Junction\_Detail]

,[Junction\_Control]

,[Vehicle\_Type]

,[Vehicle\_Manoeuvre]

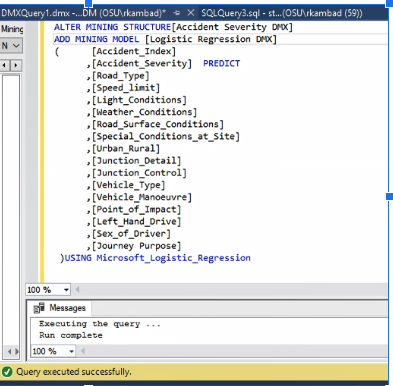
,[Point\_of\_Impact]

,[Left\_Hand\_Drive]

,[Sex\_of\_Driver]

,[Journey Purpose]

)USING Microsoft\_Logistic\_Regression  
In the below screenshot, we see the Logistic Regression DMX mining model created under Accident Severity DMX mining structure.



**Model3: Neural Networks Mining Model**  
  
ALTER MINING STRUCTURE[Accident Severity DMX]  
ADD MINING MODEL [Neural Network DMX]  
( [Accident\_Index]   
 ,[Accident\_Severity] PREDICT

,[Road\_Type]

,[Speed\_limit]

,[Light\_Conditions]

,[Weather\_Conditions]

,[Road\_Surface\_Conditions]

,[Special\_Conditions\_at\_Site]

,[Urban\_Rural]

,[Junction\_Detail]

,[Junction\_Control]

,[Vehicle\_Type]

,[Vehicle\_Manoeuvre]

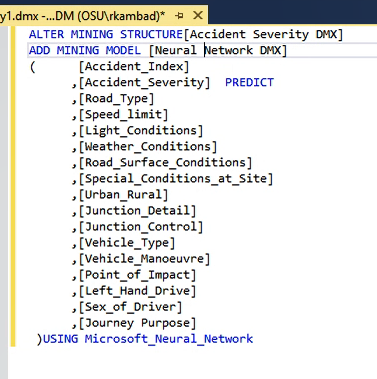
,[Point\_of\_Impact]

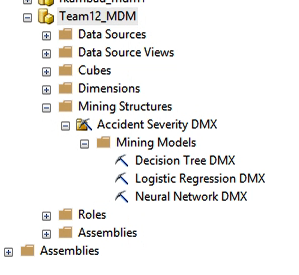
,[Left\_Hand\_Drive]

,[Sex\_of\_Driver]

,[Journey Purpose]

)USING Microsoft\_Neural\_Network



The below screenshot shows all the three mining models we created under the Accident Severity mining structure.  
  


INSERT INTO MINING STRUCTURE [Accident Severity DMX]

( [Accident\_Index]

,[Accident\_Severity]

,[Road\_Type]

,[Speed\_limit]

,[Light\_Conditions]

,[Weather\_Conditions]

,[Road\_Surface\_Conditions]

,[Special\_Conditions\_at\_Site]

,[Urban\_Rural]

,[Junction\_Detail]

,[Junction\_Control]

,[Vehicle\_Type]

,[Vehicle\_Manoeuvre]

,[Point\_of\_Impact]

,[Left\_Hand\_Drive]

,[Sex\_of\_Driver]

,[Journey Purpose]

) OPENQUERY ([UK Accidents Database],

‘SELECT TOP 100000

DimAccidents.Accident\_Index

,Accident\_Severity

,Road\_Type

,Speed\_limit

,Light\_Conditions

,Weather\_Conditions

,Road\_Surface\_Conditions

,Special\_Conditions\_at\_Site

,Urban\_Rural

,Junction\_Detail

,Junction\_Control

,Vehicle\_Type

,Vehicle\_Manoeuvre

,Point\_of\_Impact

,Left\_Hand\_Drive

,Sex\_of\_Driver

,Journey\_Purpose

FROM DimAccidents ,

DimVehicles

WHERE DimAccidents.Accident\_Index = DimVehicles.Accident\_Index

AND (Road\_Type <> -1 OR Road\_Type != 9)

AND Light\_Conditions != -1

AND Weather\_Conditions != -1

AND Road\_Surface\_Conditions != -1

AND Special\_Conditions\_at\_Site != -1

AND Junction\_Detail ! = -1

AND Junction\_Control ! = -1

AND Vehicle\_Type != -1

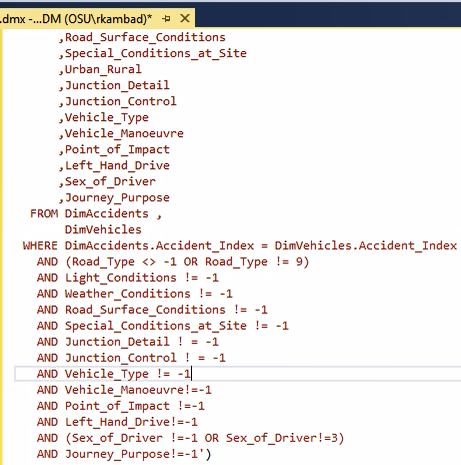
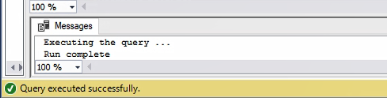
AND Vehicle\_Manoeuvre!=-1

AND Point\_of\_Impact !=-1

AND Left\_Hand\_Drive!=-1

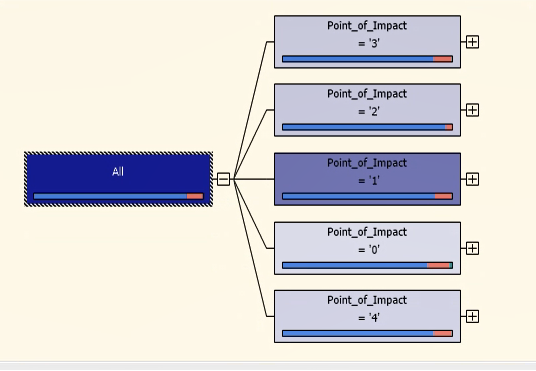
AND (Sex\_of\_Driver !=-1 OR Sex\_of\_Driver!=3)

AND Journey\_Purpose!=-1’)

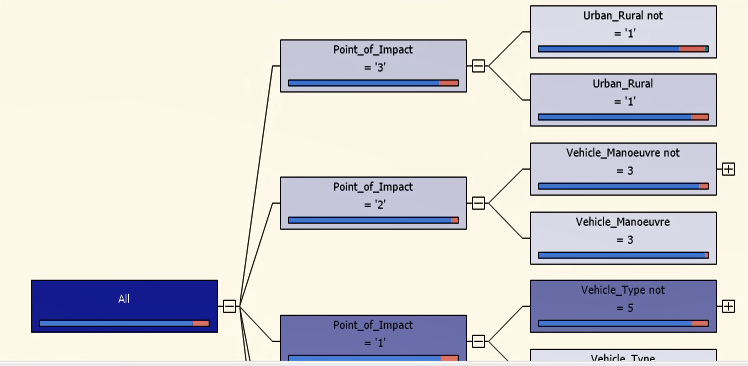
  
  


**Model 1: Decision Tree**

1st Level

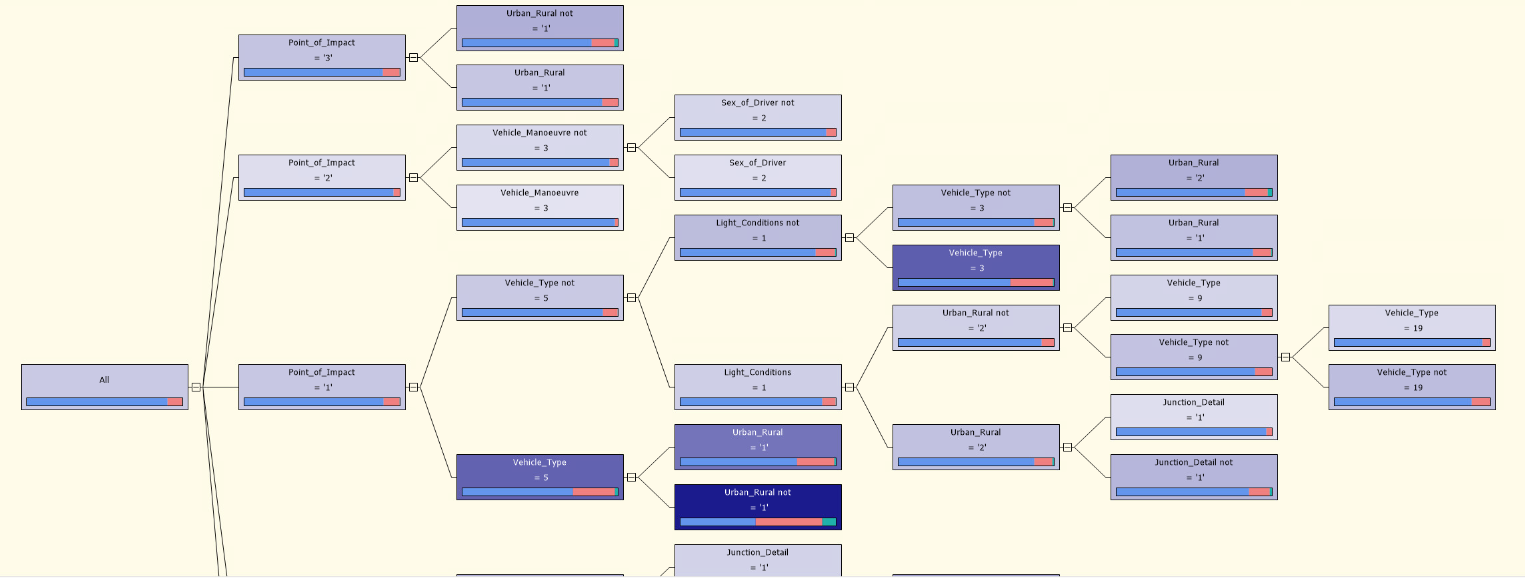
  
From the above decision tree output, we see that whether the point of impact is front, rear, left or right, it is the most important factor in predicting accident severity.

2nd Level

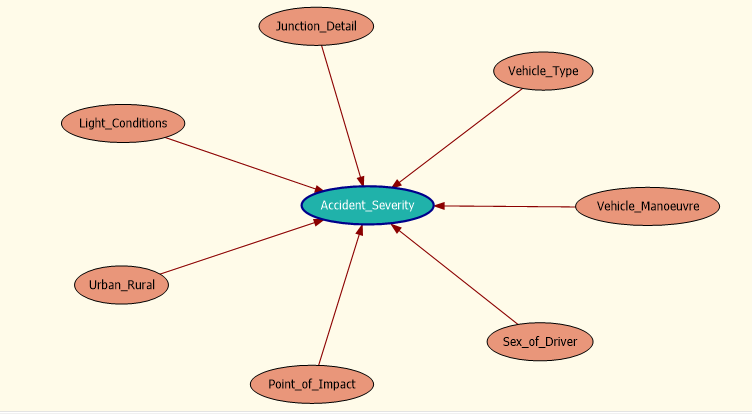
  
  
Depending upon the point of impact,

* If the point of impact is 3, the accident severity depends on location if it is urban or rural.
* If the point of impact is 2, the accident severity depends on the vehicle manoeuvre or not.
* If the point of impact is 1, the accident severity depends on the vehicle type.

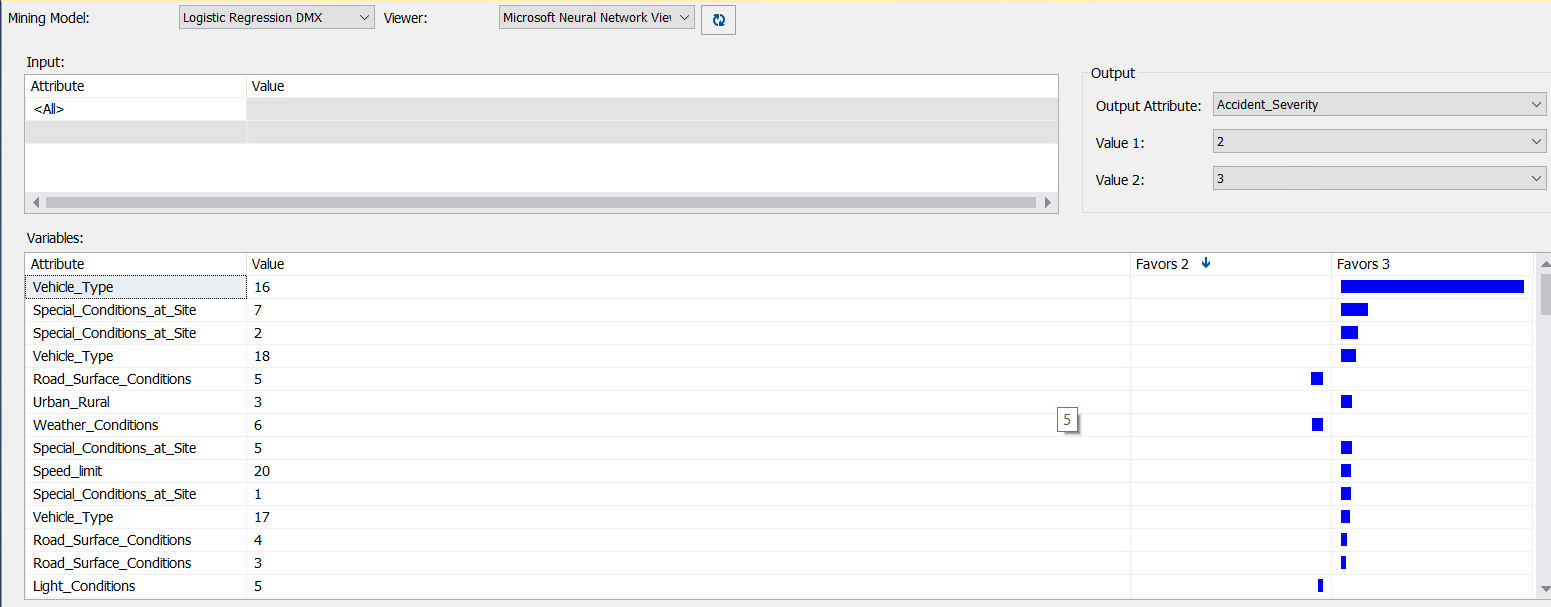


**Complete Decision Tree**  


**Dependency Network**

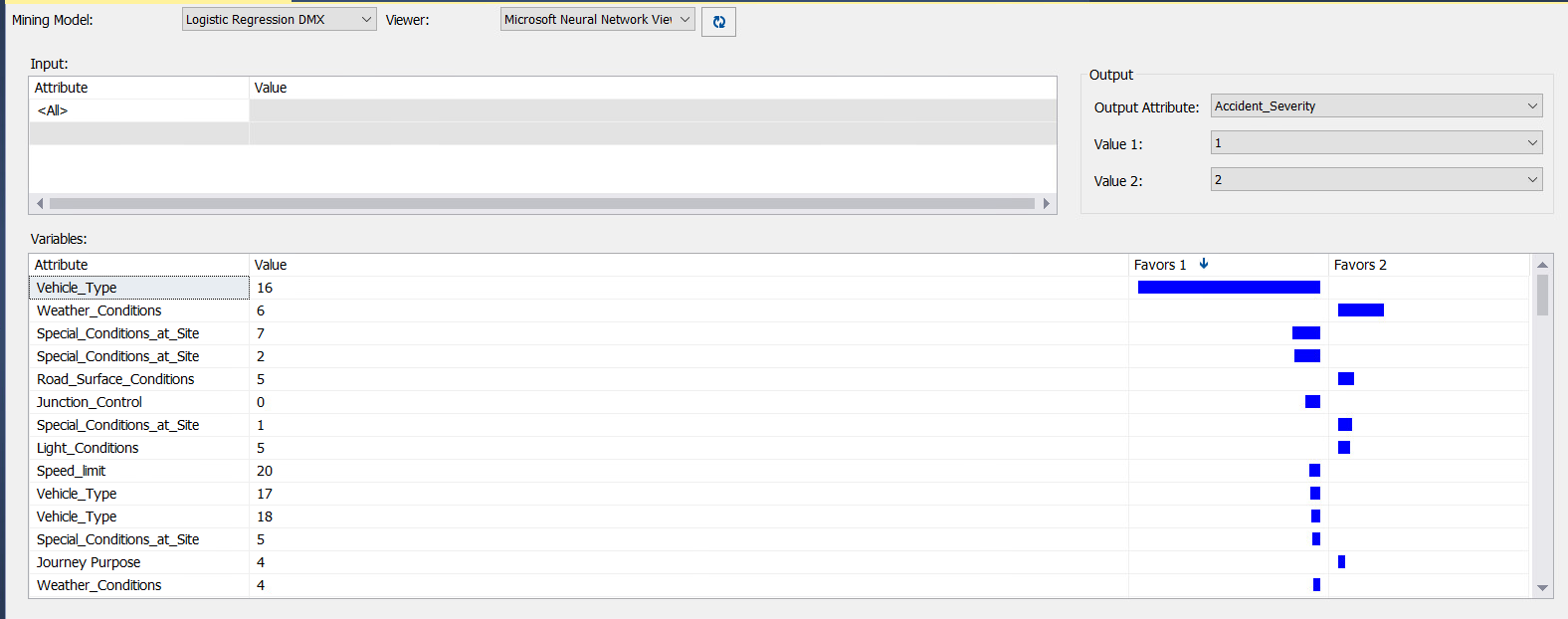
  
From the above screenshots of the complete decision tree and dependency network, we conclude that point of impact, location whether urban\_rural ,light conditions, junction detail, vehicle type, vehicle manoeuvre, sex of the driver play good roles in predicting accident severity.

**Model 2: Logistic Regression**



From the above logistic regression output we can say that, when we are considering Accident Severity 2 and 3, Vehicle\_Type value =16 favours Accident Severity= 3. Special\_conditions\_at\_site=7 is the next important variable which favours Accident Severity 3 the most.

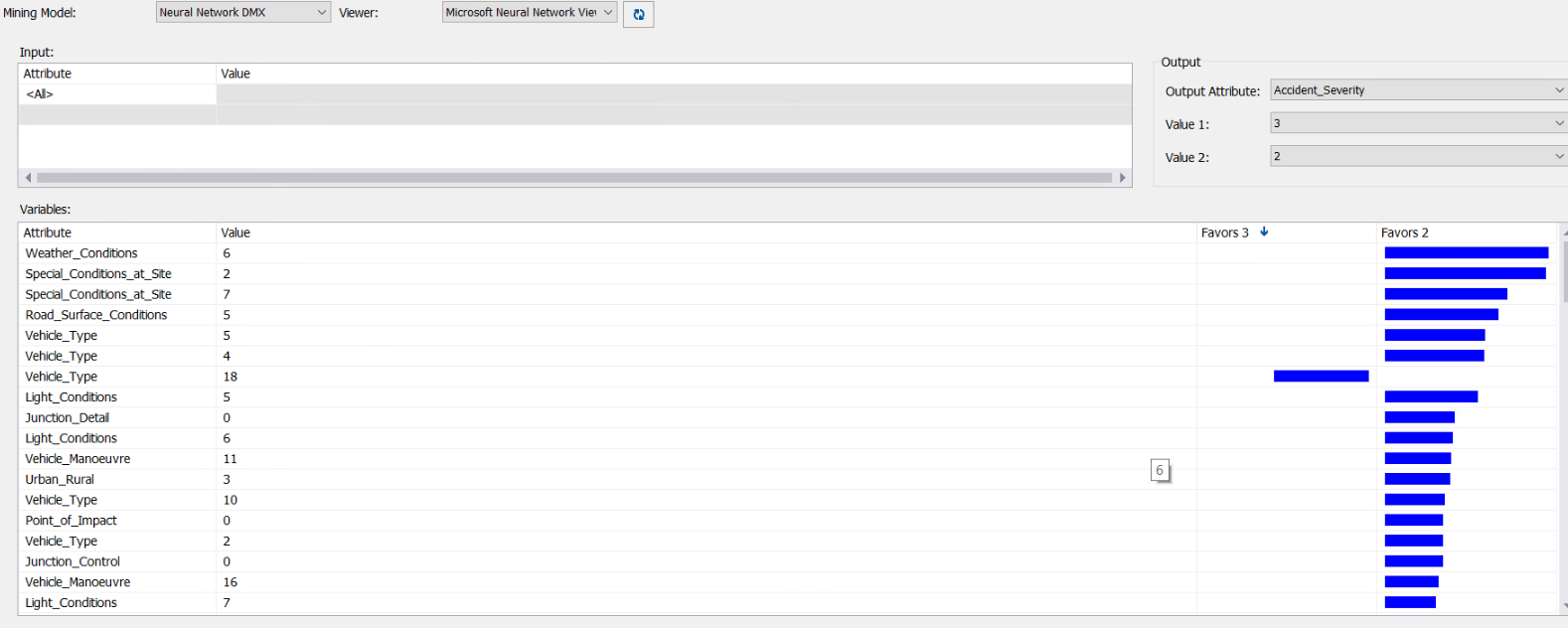
In similar manner when we are considering Accident Severity 1 and 2, vehicle\_Type value 16 favors Accident Severity 1 and Weather\_Conditions = 6 favours Accident Severity 2 the most.



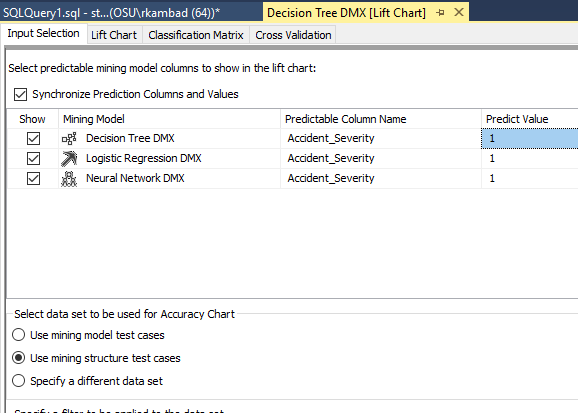
**Model3: Neural Network**

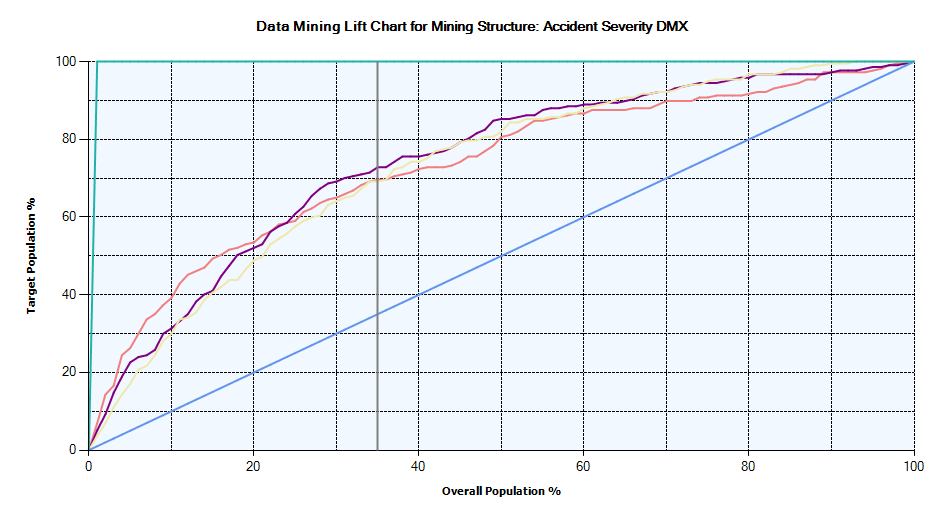
The Microsoft Neural Network algorithm combines each possible state of the input attribute with each possible state of the predictable attribute and uses the training data to calculate probabilities.

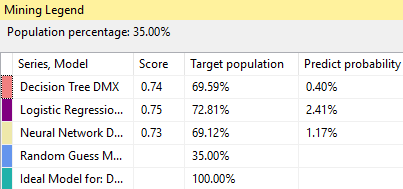
From the below neural network output we can say that weather condition = 6 favors the Accident\_Severity= 2 the most. Also, 2nd most important value is special conditions at site = 2, it favors accident severity 2.



**Comparison of the results**   
  
**Lift Chart**  
Lift chart helps us to visualize the improvement we get when we use a data mining model when compared to the random guess model. Using the mining structure test cases, we constructed the lift chart for all the 3 mining models for accident severity = 1 (Fatal)



  
The y-axis is the accuracy measure for the corresponding population percentage we targeted which is 35%(vertical grey line in the above screenshot ).



From the above charts, we can say that if we target 35% of the population,

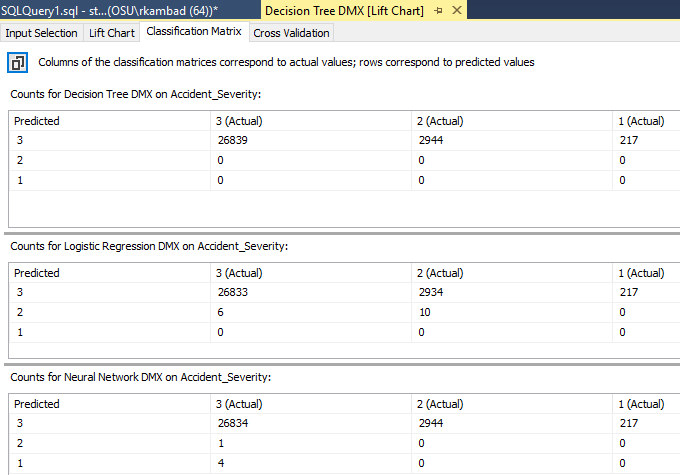
* The Random Guess Model will correctly identify 35% of all the accident severity =1(fatal) within the population
* The ideal line/model for decision tree will correctly identify 100% of all the accident severity =1(fatal) within the population. (slope=1)
* The Decision Tree model will correctly identify 69.6% of all the accident severity =1(fatal) within the population
* The Logistic Regression model will correctly identify 72.81% of all the accident severity =1(fatal) within the population
* The Neural Network model will correctly identify 69.12% of all the accident severity =1(fatal) within the population

The lift score is highest for Logistic Regression model. (Although, there is not much difference between all the 3 models).

Interpreting Predict probability

* To identify the accidents from the Logistic Regression model which have accident severity = 1(fatal), we need to use a query to retrieve cases with a predict probability of at least 2.41%
* To identify the accidents from Decision Tree model which have accident severity = 1(fatal), we need to use a query to retrieve cases with a predict probability of at least 0.40%
* To identify the accidents from Neural Network model which have accident severity = 1(fatal), we need to use a query to retrieve cases with a predict probability of at least 1.71%

(Reference for lift score, chart, predict probability: Lecture 7 slide 39)

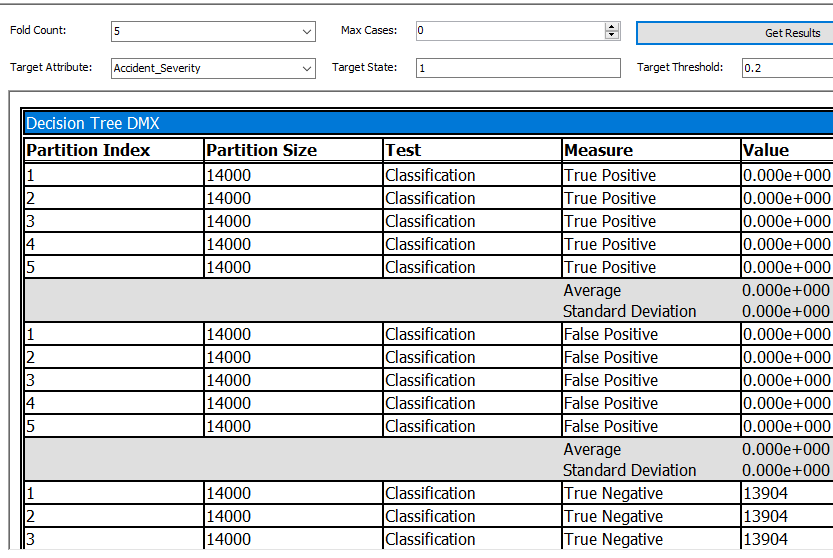
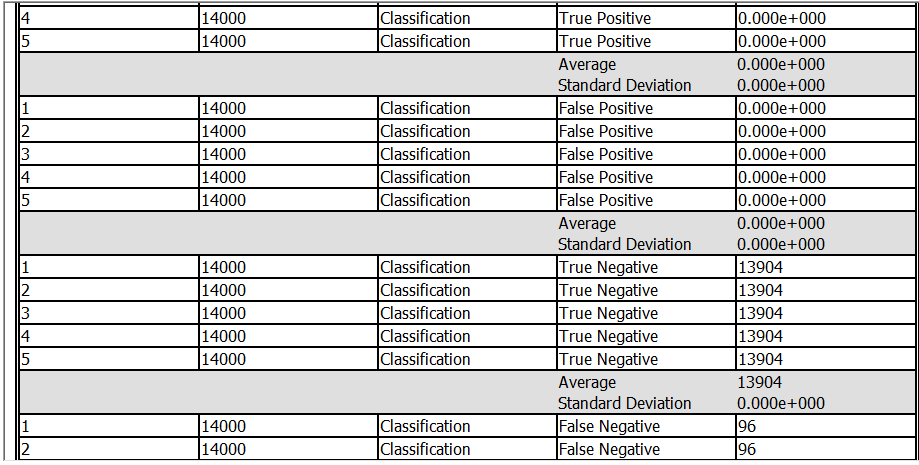
**Classification Matrix**  
  


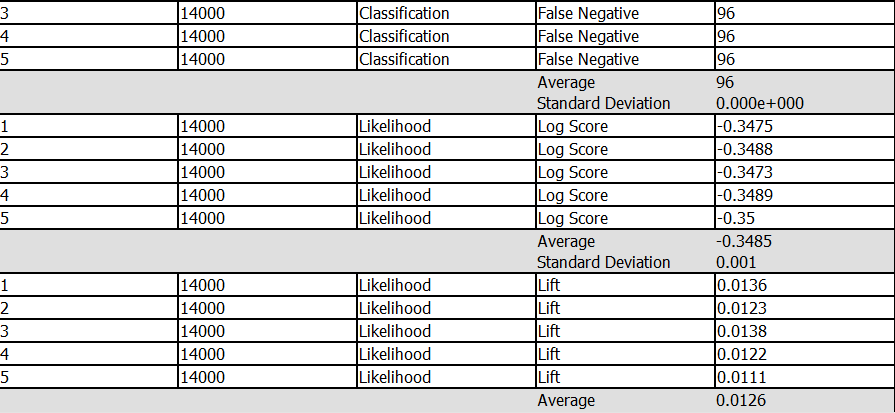
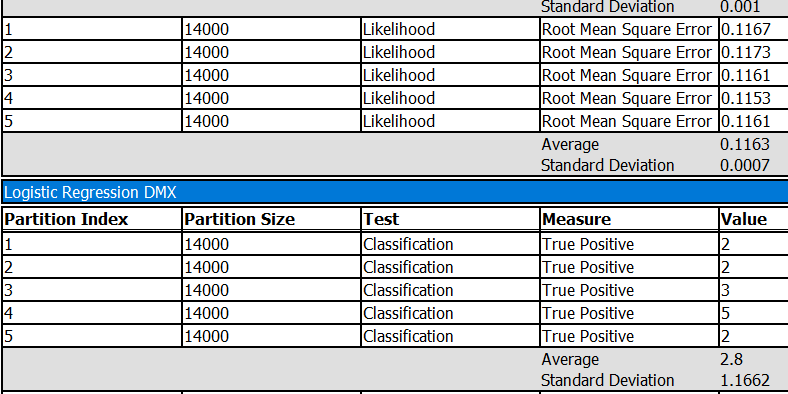
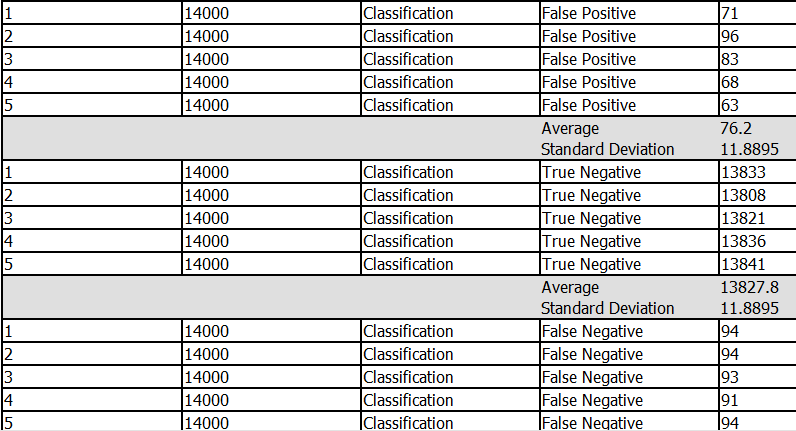
**Classification interpretations**

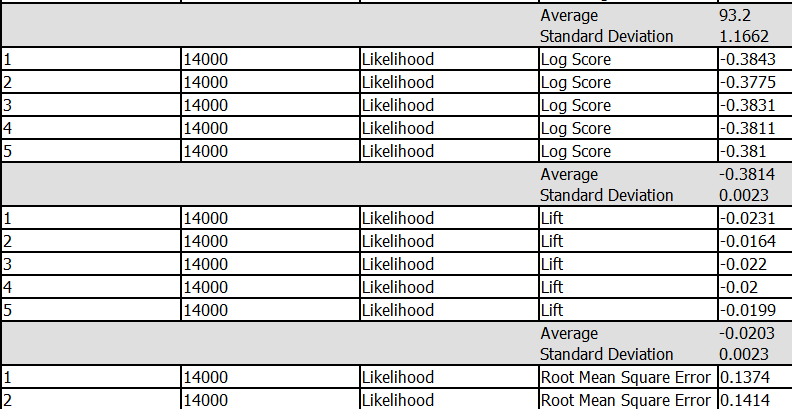
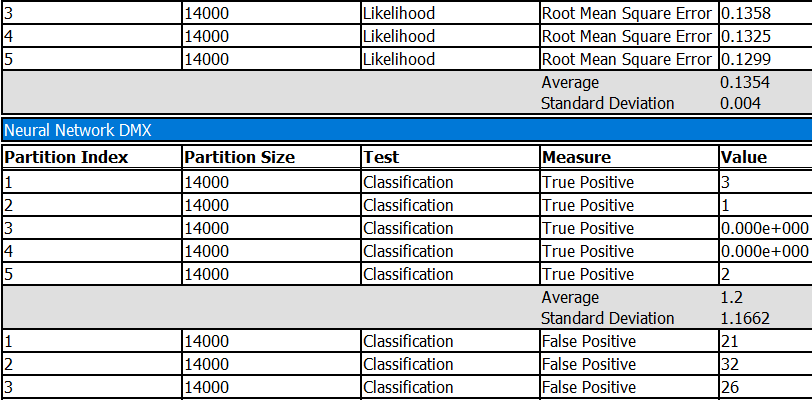
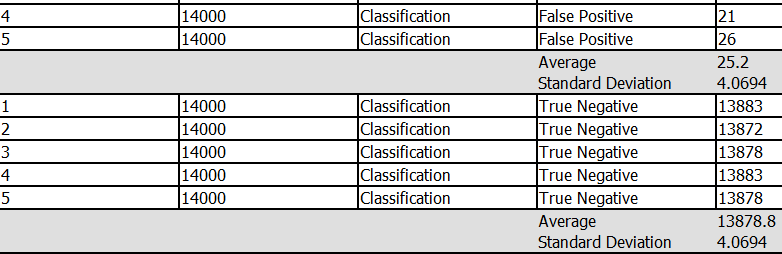
Decision tree model results predicted that 26,839 accidents would have had accident severity level 3 correctly. 2,944 accidents who has severity level 3, predicted wrongly that it has severity level 2. And 217 accidents who has severity level 3, predicted wrongly that it has severity level 1.

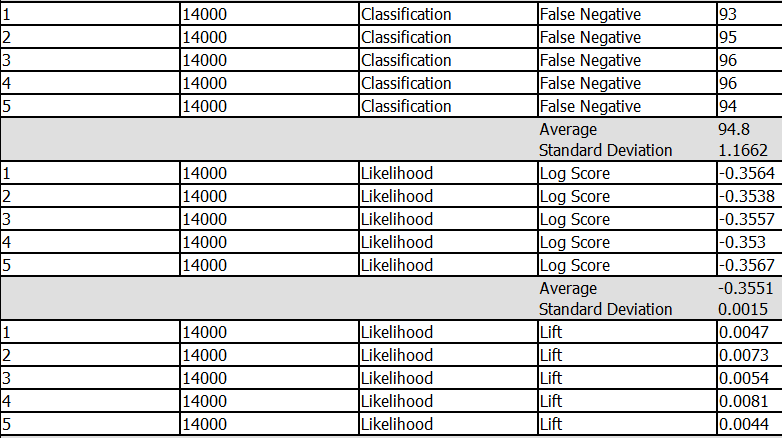
Logistic regression model results predicted that 26,833 accidents would have had accident severity level 3 correctly. 2,934 accidents who has severity level 3, predicted wrongly that it has severity level 2. And 217 accidents who has severity level 3, predicted wrongly that it has severity level 1.

Neural network model results predicted that 26,834 accidents would have had accident severity level 3 correctly. 2,944 accidents who has severity level 3, predicted wrongly that it has severity level 2. And 217 accidents who has severity level 3, predicted wrongly that it has severity level 1.

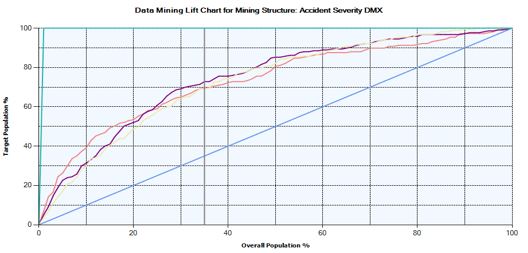
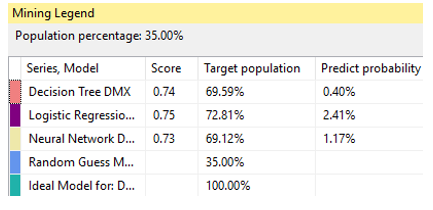
**Cross Validation**  
  
We have specified target state = 1 (accident severity = 1 or fatal) and a target threshold =0.2  
  


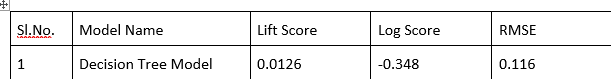
  
  



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl.No. | Model Name | Lift Score | Log Score | RMSE |
| 1 | Decision Tree Model | 0.0126 | -0.348 | 0.116 |
| 2 | Logistic Regression Model | -0.0203 | -0.381 | 0.135 |
| 3 | Neural Networks Model | 0.006 | -0.3551 | 0.135 |

**Summary of the results**  
Based on our above analysis of all the models, we conclude that decision tree is the best model.Here are the reasons to conclude so.  
  
**Lift Chart for accident severity =1**  
Based on the above screenshot after lift chart generation, we see that though the lift score for the logistic regression is good, there is not much great difference in the lift score for decision tree model. **Classification matrix**   
  
Decision tree model results predicted that 26,839 accidents would have had accident severity level 3 correctly. 2,944 accidents who has severity level 3, predicted wrongly that it has severity level 2. And 217 accidents who has severity level 3, predicted wrongly that it has severity level 1.   
  
The prediction of decision model is slightly better than the other two models.

**Cross validation**  
We have specified target state = 1 (accident severity = 1 or fatal) and a target threshold =0.2  


* The log score for the decision tree model is also closest to 0. The average prediction probability for the same model is e^(-.348) = 0.7
* The lift score indicates that there is a 1.26% improvement in the probability of the target outcome when decision tree model is used.

Since the log and lift score for decision tree is slightly good compared to the other two models, decision tree model seems to be a good model built  
  
**Conclusion**1. The decision tree model is not only more accurate than other models, but the factors explained by it are also makes sense in real world.

2. Based on whether the point of impact is front of the vehicle, rear, sides etc., we can say how severe an accident is. Also, there will be ideally a greater number of accidents in urban\_rural = 1 or 2.

3. The kind of vehicle we travel whether motorcycle 125 cc and under, light conditions and sex of the driver plays important roles in deciding accident severity.

4. Compared to the vehicle which is waiting to go (held up), the vehicles which are turning left or right or changing the lanes, they are prone to accidents.