

***Data Warehouse for Road Traffic
Accident Data in the UK***

ADVANCED INFORMATION SYSTEMS
PROJECT REPORT

Author: Tilemachos S. Doganis

Co-Authors: Spiros Kaftanis, Apostolos Kemos

INTRODUCTION

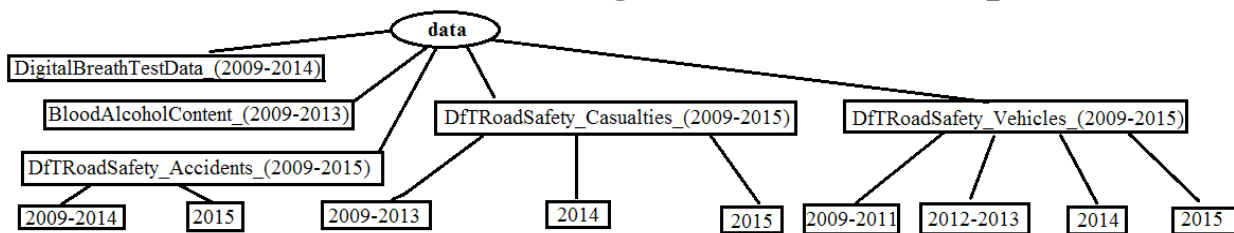
The present project implements a Data Warehouse. This system draws data from files on the UK government website, regarding traffic accidents ¹, and consists of two basic components: The ETL (extract, transform, load) input structure and the hypercube (OLAP Cube) which is used for organization and analysis of the data through the corresponding queries.

The following software was used for this project:

- Microsoft SQL Server 2014 (Developer Edition)
- SQL Server Data Tools for Visual Studio 2015
- Microsoft Excel 2010
- PowerBI (2.51.4885.543)

A technical description of the basic components is next, followed by an explanation for the decisions taken during their design, with examples from the cube visualization at the end.

Instructions for usage on another computer



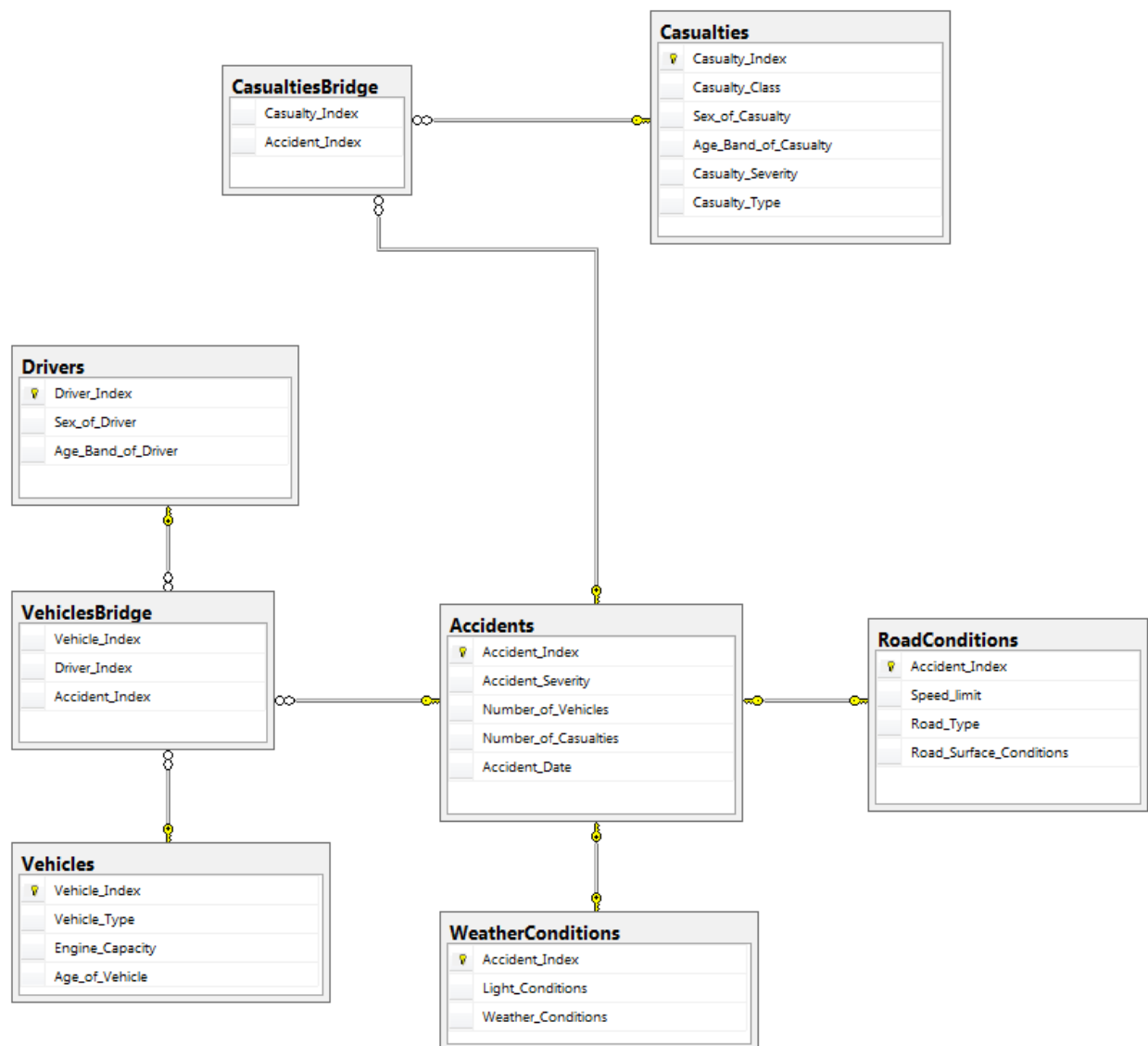
1. For the ETL to load the data, they must be structured as follows: The .csv files of Accidents are grouped together, while the Casualties and Vehicles files are placed in separate folders according to the year.
2. Furthermore, for the Breath Test Data, Blood Alcohol Content, Casualties 2009-2013, Vehicles 2009-2011 and Vehicles 2012-2013 the positions of the corresponding folders must be defined on the Collection Tab of the Foreach Loop Components, while for the rest of the years the filepaths of the corresponding .csv files must be defined in their Flat File Managers.
3. **Note:** The ETL package, to function properly, requires that the corresponding tables do not already exist in the database (otherwise they first have to be manually removed).

In addition, the Delay Validation property of the Data Flow Tasks has been set to 'True', in order to avoid errors in I/O operations to and from tables not yet created . If checking their Data Flows is desired before the Create Tables Component has been run, an error will appear in the package, which will eventually disappear with the creation of the tables.

¹Source: <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>

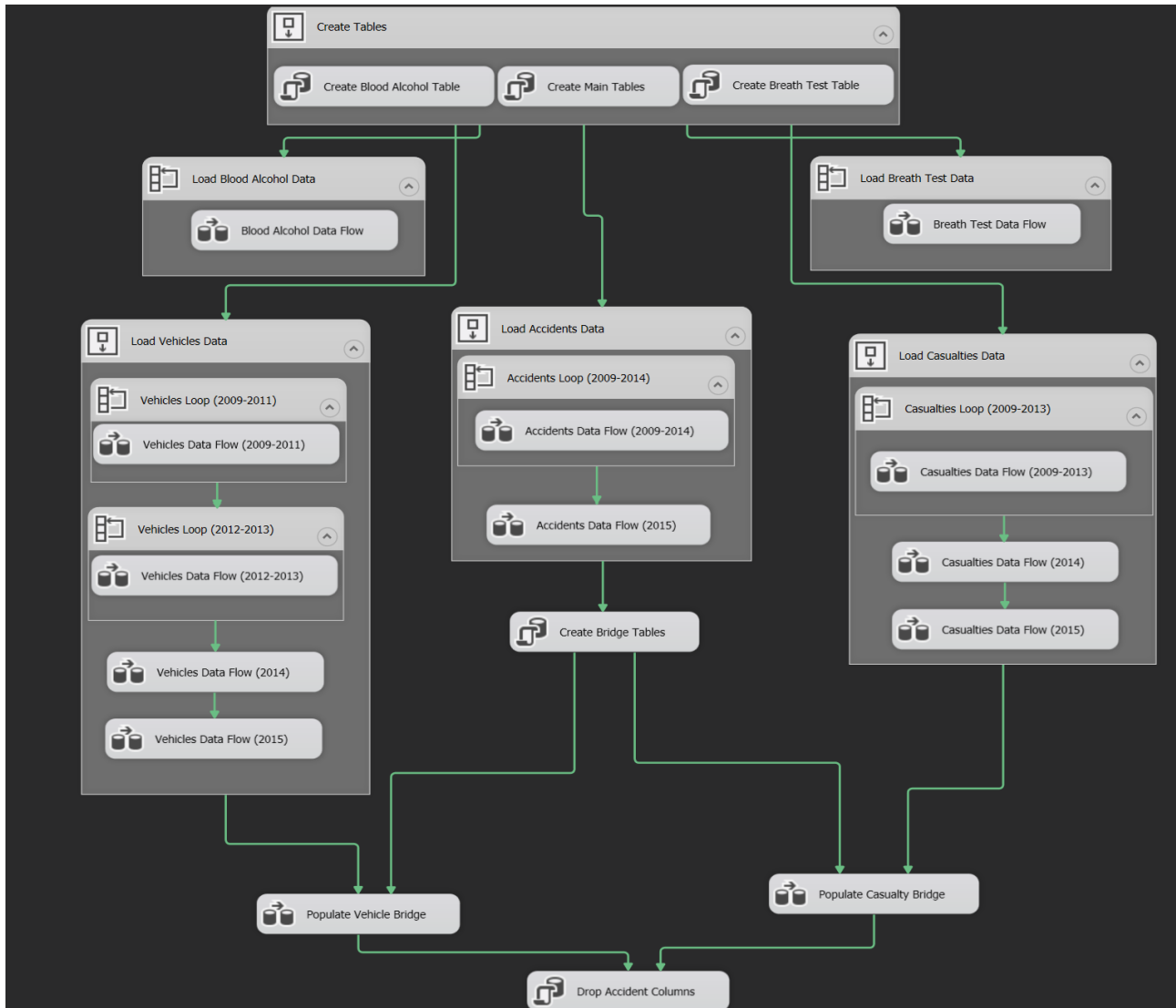
Star Schema

From the available columns of the three basic sources Accidents, Casualties and Vehicles, six tables are produced: From Vehicles the 'Vehicles' and 'Drivers' tables, from Accidents the 'Accidents', 'RoadConditions' and 'WeatherConditions' tables, and from Casualties the 'Casualties' table. The Accidents table has been placed in the center as a Fact table. For the Vehicles, Drivers and Casualties tables, as they comprise separate entities, manually produced Indexes are necessary. Because they are connected to the Accidents table through Many-to-One relationships though, Bridge Tables are used for their connections with Accidents. (The use of Accident_Index as a Foreign Key was also tried, but in SSAS it was only possible in One-to-One assignment)



ETL – Control Flow

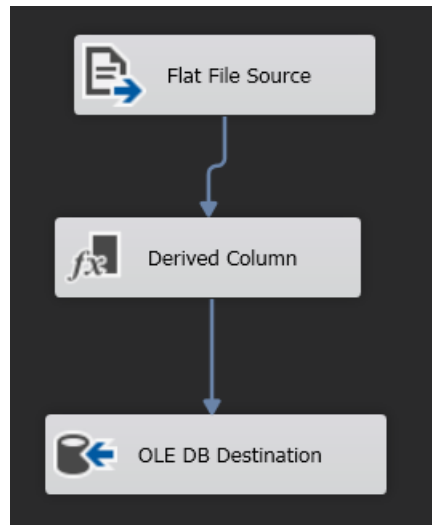
ETL is implemented through the ETL.dtsx package in the SSIS project and its Control Flow is seen below:



Initially, the database tables that will store the data are created, and afterwards the data are loaded from the corresponding files of .csv form. Because some years of Vehicles and Casualties have slightly different column names, while in the more recent ones additional columns exist as well, data loading takes place in year groups with common columns, which are placed accordingly in subfolders. Loops have been implemented using Variables (one for each table). Lastly, Indexes are placed on the Bridge Tables to connect them with Accidents and the Accident_Index columns are deleted from the corresponding tables, since the connection is now made through the Bridge Tables

ETL - Data Flow

Data Flows have the following form:



- Flat File Source is linked via a Connection to the corresponding .csv file (or a loop iterator). During the assignment of the csv file's columns with the output columns, restricted characters are replaced (dashes, semicolons, parentheses) to avoid potential problems in their later use. Furthermore, the on the Error Output option the “Ignore Failure” field has been selected in the error column for the attributes that can take Null values, so that if a problematic value is given, outside of the accepted value range, it will be automatically replaced with Null.
- Derived Column filters the columns and, via an Expression, replaces any "-1" value with Null, which is the default value of SSDT for problematic / missing values.
- OLE DB Destination loads the data into the DataWarehouseDB database, via the “localhost.DataWarehouseDB” Connection. Because of the differences in names and number of columns per year in every one of the three datasets, the non-existing columns are ignored so that they will be automatically filled with Null in the database, while potentially different column names are matched to an arbitrarily chosen common name.
- Blood Alcohol Data Flow has two additional components: a Conditional Split to reject the rows where Blood Alcohol Content is 'Unknown' (unnecessary data) and a Data Conversion to convert the remaining values to two-byte integer numeric values.

OLAP Cube

For the creation of the hypercube, connection to the Analysis Services (SSAS) of SSDT is required. Specifically, after the database DataWarehouseDB is defined as a Data Source via its corresponding Connection (localhost.DataWarehouseDB), a Data Source View of this source is created.

The Blood Alcohol and Breath Test tables have no Accident Index, and cannot therefore be combined with the rest of the tables, but they can be used independently for drawing conclusions.

Subsequently, Named Calculations are defined for the database tables, which assign text to the numbers that, until now, represented the categorical data (age, sex, etc.). This is only done to the categorical data used in this project, but they can be defined for every required attribute in a similar manner, if necessary. To implement these assignments, the corresponding Named Calculation is chosen as a Name Column in the Properties of each categorical attribute.

Before constructing the cube, its dimensions are defined from subsets of the table columns, while an artificial time dimensions is produced which contains all dates and months for the years 2009-2015.

Lastly, the hypercube is constructed, using the Accidents table as a Measure Group and the two Bridge Tables as hidden Intermediate Measure Groups, to implement the connection between the other tables and Accidents. Afterwards, the Accidents table is added as a dimension, along with the dimension tables, while the time dimension is added through a Regular Relationship with the Measure Group, in order for the cube to support time-related queries. Furthermore, the two dimensions of the secondary tables are added along with their Count Measures.

Named Sets

Named sets are Multidimensional Expression (MDX) scripts, which can be stored in the cube and provide with a way to restrict the cube to a specific subregion of interest so that any future slicing will be possible in that sub-cube.


We created two named sets, one which returns the three months with the most casualties:


```
CREATE DYNAMIC SET CURRENTCUBE.[Top 3 Years]
AS TopCount
(
    [DimTime].[Year].[Year].MEMBERS,
    3,
    [Measures].[Number Of Casualties]
), DISPLAY_FOLDER = 'Year Set' ;
```

and is used in the Pivot Table as seen below:


	A	B	C	D
1		Column Labels		
2		Calendar 2009	Calendar 2010	Calendar 2011
3	Number Of Casualties	222146	208648	203950


Drag fields between areas below:

 Report Filter

 Column Labels

Top 3 Years ▼

 Row Labels

 Values

Number Of Casualties ▼

☐ Defer Layout Update

Update

and the:

```
CREATE DYNAMIC SET CURRENTCUBE.[Top 3 Fatal Years]
AS TopCount
(
    [DimTime].[Year].[Year].MEMBERS,
    3,
    [Measures].[Fatalities]
), DISPLAY_FOLDER = 'Year Subset' ;
```

which returns the three years with the most deaths in fatal accidents.

	A	B	C	D
1		Column Labels		
2		Calendar 2009	Calendar 2010	Calendar 2011
3	Fatalities	2,786	2,383	2,372

Drag fields between areas below:

Report Filter

Column Labels: Top 3 Fatal Years

Row Labels

Σ Values: Fatalities

Defer Layout Update

Update

Calculated Members

Calculated members are new Measures that emerge from some numeric combination of existing Measures. Their definitions are stored in the cube, but their results are calculated at the moment of querying.

We created a calculated member which calculates which percentage of total casualties is fatal, dividing the Fatalities measure with the Number of Casualties measure. THE Pivot Table σε combination with the Year – Month – Date hierarchy can be seen below:

Row Labels	Calculated Average Member	Choose fields to add to report:
+ Calendar 2009	1.36	<input type="checkbox"/> Weather Conditions
+ Calendar 2010	1.35	+ <input type="checkbox"/> Severity
- Calendar 2011	1.35	<input type="checkbox"/> DimCasualties
+ January 2011	1.34	+ <input type="checkbox"/> Accident - Casualty
+ February 2011	1.34	+ <input type="checkbox"/> More fields
+ March 2011	1.33	<input type="checkbox"/> DimTime
+ April 2011	1.37	+ <input checked="" type="checkbox"/> Year - Month - Date
+ May 2011	1.35	+ <input type="checkbox"/> Month Set
+ June 2011	1.36	+ <input type="checkbox"/> Top 3 Years
+ July 2011	1.35	+ <input type="checkbox"/> More fields
+ August 2011	1.38	<input type="checkbox"/> DimVehicles
+ September 2011	1.34	+ <input type="checkbox"/> Accident - Vehicle
+ October 2011	1.34	+ <input type="checkbox"/> More fields
+ November 2011	1.32	
+ December 2011	1.34	
+ Calendar 2012	1.34	
+ Calendar 2013	1.32	
- Calendar 2014	1.33	
+ January 2014	1.31	
+ February 2014	1.32	
+ March 2014	1.33	
+ April 2014	1.35	
+ May 2014	1.34	
+ June 2014	1.33	
+ July 2014	1.34	
+ August 2014	1.37	
+ September 2014	1.31	
+ October 2014	1.32	
+ November 2014	1.31	
+ December 2014	1.33	
+ Calendar 2015	1.33	
Grand Total	1.34	

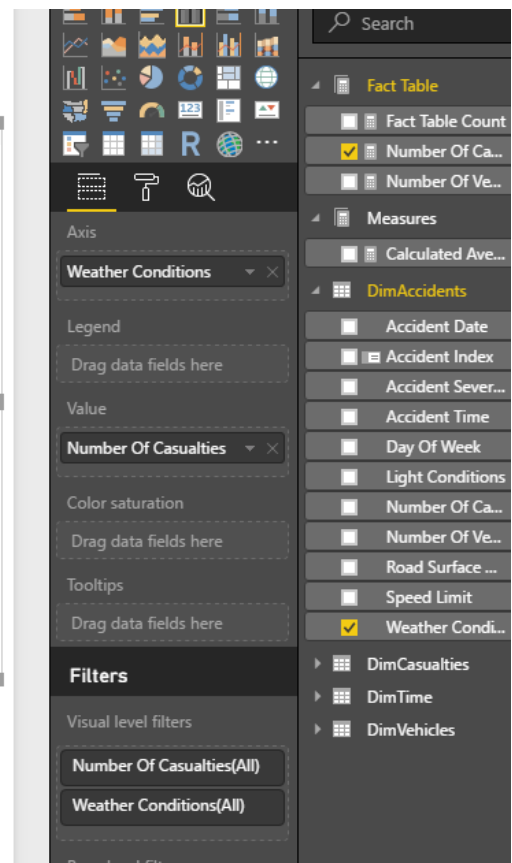
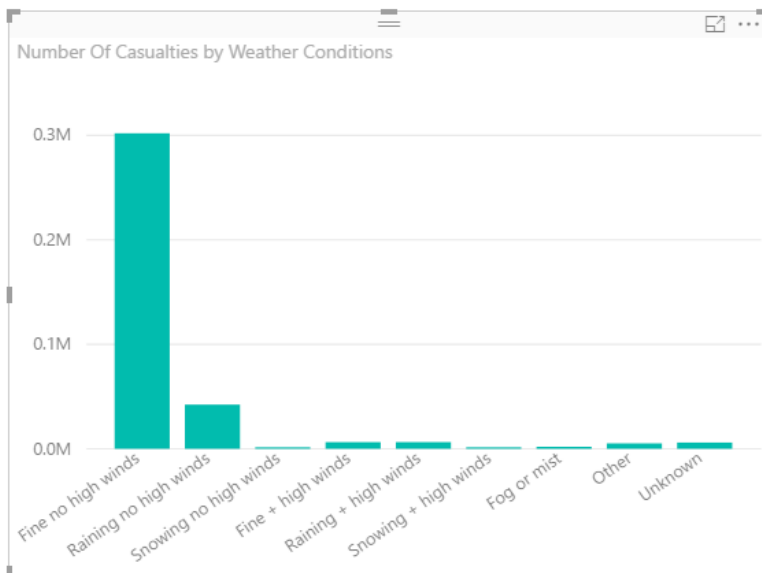
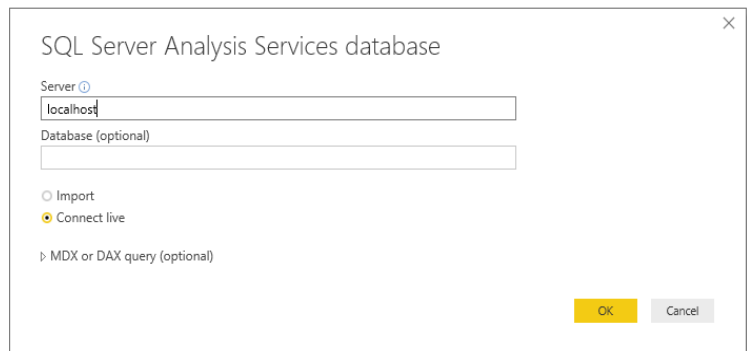
Furthermore, we created another Calculated Member which finds fatal casualties: (through IIF and ISEMPTY, SSAS is prevented from automatically converting 0 to Null)

```
CREATE MEMBER CURRENTCUBE.[Measures].[Fatalities]
AS IIF(
    ISEMPTY([Measures].[Number of Casualties],[Casualties].[Casualty Severity].&[1]))
    ,0
    ,([Measures].[Number of Casualties],[Casualties].[Casualty Severity].&[1])
),
VISIBLE = 1 , ASSOCIATED_MEASURE_GROUP = 'Accidents';
```

PowerBI Visualization

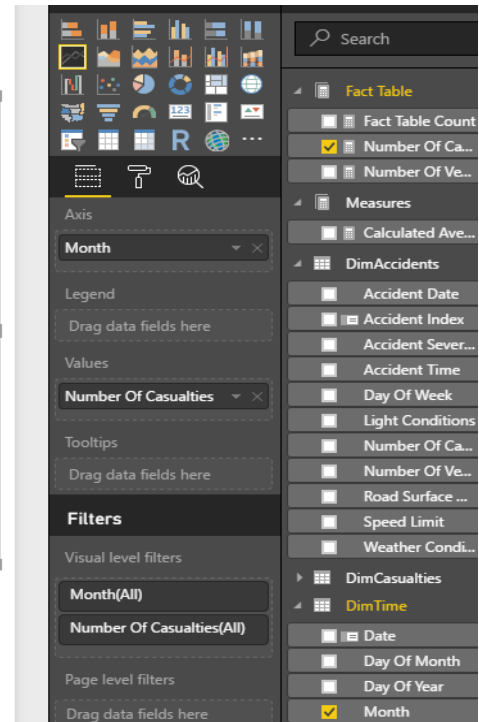
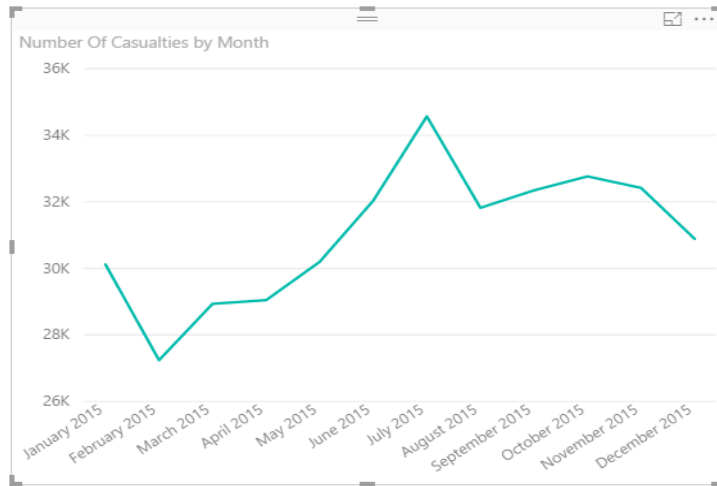
For visualization we connected to Analysis Services through PowerBI to access the data from the cube.

We plotted the histogram of totalcasualties according to weather conditions to conclude under which ones most accidents occurred.



It is observed that most accidents happen under good weather conditions, with accidents during rainfall at the second position. The conclusion can therefore be drawn that rain has some contribution to accidents, albeit not a very large one. Wind on the other hand seems to be uncorrelated to the number of accidents.

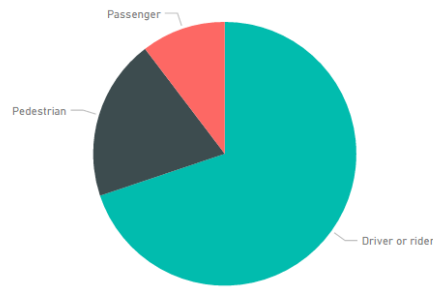
Afterwards we plot a line chart of a specific year's months (2015) and the number of casualties to examine the trend in that year.



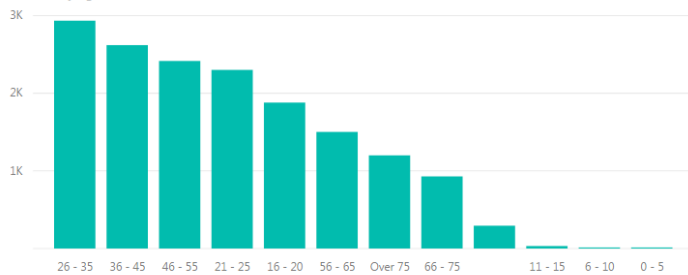
We observe that there is an increase that peaks at July, which fades at Autumn and the following months.

Fatal traffic accidents

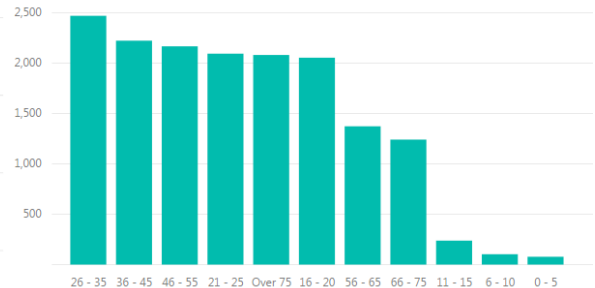
Fatalities by Casualty Class



Fatalities by Age Band Of Driver



Fatalities by Age Band Of Casualty



In a similar manner some graphs are created, related to the examination of fatal casualties. In the first graph, one can see that the majority are the drivers. Similarly, it seems that most deaths correspond with young to middle age groups, decreasing in younger and older ages. Possibly, this is observed because in younger ages less people drive and those that do, probably are more careful due to lack of experience. In older age groups it stands to reason that they drive more slowly and carefully, while in middle age groups a sense of overconfidence could lead to risky driving behaviour.

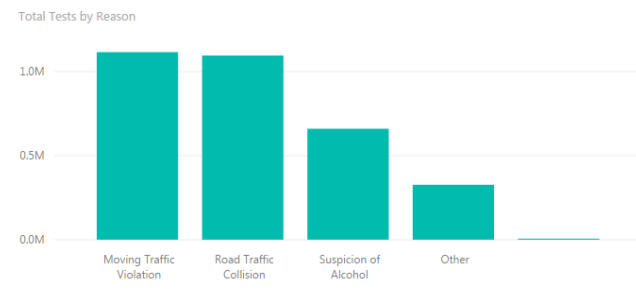
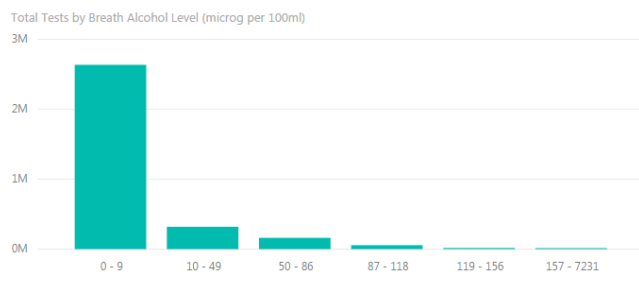
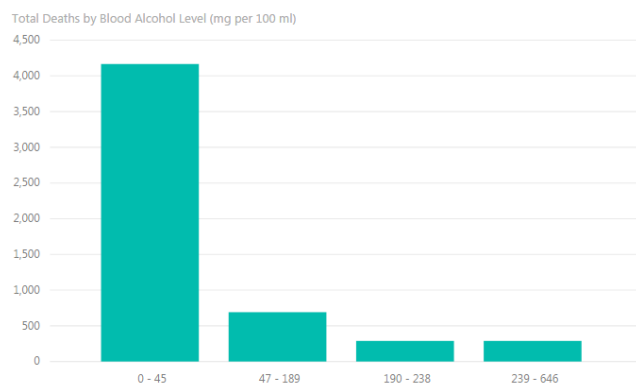
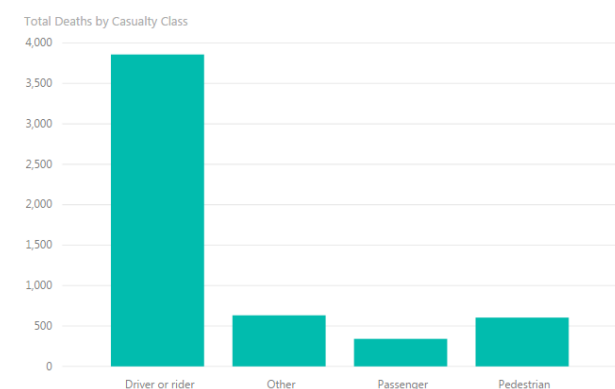
It is also worth noting that the ages of the drivers and the victims are very similar.

Alcohol-related statistics

To begin with, it is useful to mention the upper limits of allowed alcohol content in breath and blood in the UK:

Level of alcohol	England, Wales and Northern Ireland	Scotland
Micrograms per 100 millilitres of breath	35	22
Milligrammes per 100 millilitres of blood	80	50

Source: <https://www.gov.uk/drink-drive-limit>



The first conclusion is drawn from the upper left diagram, namely that the majority of autopsies were done on drivers, which is to be expected given that they are the main group that is affected by alcohol in the body. Nevertheless, the presence of deaths in the pedestrian category probably is partly comprised of people who did not pay enough attention to the street because of the influence of alcohol. From the second diagram, however, we draw the conclusion that most deaths are not due to alcohol.

In the second row of alcohol testing diagrams, it seems that even though most people are stopped due to violations, the majority has little to no alcohol in their breath. Consequently it seems that even though alcohol is connected with a number of accidents, the foremost cause is probably the human factor.

Design Decisions

- **Data choice:** The columns that were chosen were those that contained the most useful information for queries on the data warehouse, while other columns or columns with multiple Null values, like those that appear only on the most recent years, were ignored for reasons of speed and saving space.
- **Star Schema:** For the creation of the Star Schema it was decided that the Accidents table be placed as a Fact Table in the center, as it contains almost all possible measures. Certain columns were moved to distinct dimensions, such as RoadConditions and WeatherConditions, as well as Drivers which emerged from Vehicles as a separate entity.
- **Loading files with different columns:** As was mentioned in the ETL section, due to small differences in the file columns, instead of a single Foreach Loop for each one of the three main data types, they need to be split into groups with common columns and be loaded as such. In addition, because the first rows of Accidents_2015 have no equivalents in Vehicles and Casualties, they are bypassed.
- **Managing unnecessary values / columns:** In Blood_Alcohol the values “Driver” and “Driver or Rider” coexist, so the first was merged with the second through a Derived Column, while the Casualty Type column was “Fatal” everywhere, since the data originate from autopsies, so it was ignored.
- **Managing problematic values:** The '-1' value which defines the absence of an acceptable value can either remain as is, or be replaced with Null. In the first case, there is the problem that multiple non-acceptable value types exist, while because SSDT automatically converts empty fields (in string columns) and zeroes (in numeric columns) to Null, changing the '-1' values as well makes it easier to manage this type of values. For this reason the option “Retain Null Values” is selected in every Flat File Source, as well as the option “Keep Null” in every OLE DB Destination.
- **Grouping values:** Because the visualization of groups of values it more effective, wherever a grouped set is available, such as Age Band, it is chosen over Age, while in the case of Blood Alcohol Levels and Breath Alcohol Levels, grouping was done by Clusters through the corresponding property.
- **Vehicles / Casualties Primary Key dilemma:** Whether a Primary Key will be defined for Vehicles and Casualties via a Composite Key of Accident_Index and Vehicle / Casualty Reference respectively, or as a Surrogate Key via the property AUTO_INCREMENT. While for other uses the Composite Key would be more efficient, as it does not require the addition of new fields, the absence of a single, common key is problematic for the Fact_Table which requires a Foreign Key from every table. For this reason Vehicle_ID and Casualty_ID were defined as Incremental Integer columns.
- **Numeric Data Type:** Most numeric data are integers, and because the value '-1' also exists, they are defined as signed integers.
- **Categorical Data Type:** Most can either be declared as integers or as character strings. But because in the integer type, zero is regarded as Null by SSDT, to be safe they were defined as DT_STR.
- **Matching categorical values** (via data guide): One option is with the use of an SQL which would carry out the matching through the Update action on each row of the table, which would be time consuming and would increase the space required in the database by a great amount. On the other hand, it was decided that the matching would be done via Named Calculations in the Data Source View specifically for the columns used in the project.

- **Foreign Keys:** In the database it was decided not to define Foreign Keys, since they create unnecessary constraints, while the connection between tables is implemented through the Accidents table, and all the queries take place via the Hypercube and not directly on the database. The only exception are the Bridge Tables.
- **Additional Tables:** It was decided to use the additional tables concerning Breath Test and Blood Alcohol data, in order to produce a more complete image in regards to the causes and conditions of the accidents. Even though these data are not directly connected to the main ones, neither did they cover all of their timespan, it is expected that they will reflect an accurate image. Normally each of the two should be placed in a separate Cube, but because they are few and small tables it was decided that they be placed in the main one for simplicity.
- **Choosing Measures:** The Measures that are chosen have to be numeric values that represent the meaning of the events in the database. Therefore, they are selected from the Accidents, Casualties and Vehicles that already exist in the Accidents Table, as well as the Counts of Blood Alcohol and Breath Test as independent ones, because they are not connected to the other tables. Furthermore, the Measures of the Bridge Tables exist as a formal necessity and are therefore hidden.
- **Calculated Members & Named Sets:** When someone needs to reach conclusions from a database of accidents, clearly the main criteria are the number of accidents and casualties and mainly the fatal ones of the latter. For this reason the metrics Fatalities and Fatal Accidents are produced as Calculated Members to count the number of deaths and fatalities, respectively.
In addition, the Named Set 'Top 3 Years' is defined for the years with the most casualties, as well as 'Top 3 Fatal Years' and 'Fatal Accidents' for the fatality criterium.
- **Choosing Dimension Attributes:** They are chosen based on which factors are expected to contribute the most in accidents, or which ones better describe the subjects. Usually they are age, sex, position during the accident and type of injury, for people, while regarded as contributing factors are the road conditions, weather conditions, the presence of alcohol, and once again, age. The selected attributes can be seen in the Star Schema on Page 3.